MICROBIAL GENOMICS

RESEARCH ARTICLE

Reis and Cunha, *Microbial Genomics* 2021;7:000664

DOI 10.1099/mgen.0.000664

The open pan-genome architecture and virulence landscape of *Mycobacterium bovis*

Ana C. Reis[1,2] and Mónica V. Cunha[1,2,*]

Abstract

Animal tuberculosis (TB) is an emergent disease caused by *Mycobacterium bovis*, one of the animal-adapted ecotypes of the *Mycobacterium tuberculosis* complex (MTC). In this work, whole-genome comparative analyses of 70 *M. bovis* were performed to gain insights into the pan-genome architecture. The comparison across *M. bovis* predicted genome composition enabled clustering into the core- and accessory-genome components, with 2736 CDS for the former, while the accessory moiety included 3897 CDS, of which 2656 are restricted to one/two genomes only. These analyses predicted an open pan-genome architecture, with an average of 32 CDS added by each genome and show the diversification of discrete *M. bovis* subpopulations supported by both core- and accessory-genome components. The functional annotation of the pan-genome classified each CDS into one or several COG (Clusters of Orthologous Groups) categories, revealing 'transcription' (total average CDSs, *n*=258), 'lipid metabolism and transport' (*n*=242), 'energy production and conversion' (*n*=214) and 'unknown function' (*n*=876) as the most represented. The closer analysis of polymorphisms in virulence-related genes in a restrict group of *M. bovis* from a multi-host system enabled the identification of clade-monomorphic non-synonymous SNPs, illustrating clade-specific virulence landscapes and correlating with disease severity. This first comparative pan-genome study of a diverse collection of *M. bovis* encompassing all clonal complexes indicates a high percentage of accessory genes and denotes an open, dynamic non-conservative pan-genome structure, with high evolutionary potential, defying the canons of MTC biology. Furthermore, it shows that *M. bovis* can shape its virulence repertoire, either by acquisition and loss of genes or by SNP-based diversification, likely towards host immune evasion, adaptation and persistence.

DATA SUMMARY

The newly sequenced data included in this work are deposited under the following Biosample accession numbers: SAMN17004141–SAMN17004143, SAMN17004145–SAMN17004174, SAMN17004176–SAMN17004184 and under the Bioproject accession number PRJNA682618 at a public domain server in the National Centre for Biotechnology Information (NCBI) SRA database.

All supporting data, code and protocols have been provided within the article or through supplementary data files which can be found at https://doi.org/10.6084/m9.figshare.15025965.

INTRODUCTION

The *Mycobacterium tuberculosis* complex (MTC) encompasses 12 closely related mycobacteria that cause tuberculosis (TB) in a variety of mammalian hosts [1]. *M. tuberculosis* is the most well-known member, as it is a prominent human pathogen, accounting for approximately 10 million new cases and 1.4 million deaths in 2018 [2]. *Mycobacterium bovis* is the ecotype

Received 22 June 2021; Accepted 03 August 2021; Published 29 October 2021

Author affiliations: [1]Centre for Ecology, Evolution and Environmental Changes (cE3c), Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal; [2]Biosystems & Integrative Sciences Institute (BioISI), Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal.

***Correspondence:** Mónica V. Cunha, mscunha@fc.ul.pt

Keywords: core-genome; *Mycobacterium bovis*; pan-genome; virulence; WGS; comparative genomics.

Abbreviations: Af1, African; Af2, African 2; BCG, bacillus Calmette-Guérin; CDS, protein-coding sequence; COG, clusters of orthologous groups; COGtriangles, cluster of orthologous groups triangles; DG, distance-guide; ESX, specialized secretion system; Eu1, European 1; Eu2, European 2; Eu3, European 3; GATK, genome analysis toolkit; GTR, general time reversible; IGV, integrated genomics viewer; MEGA, molecular evolutionary genetics analysis; ML, maximum likelihood; MTC, *Mycobacterium tuberculosis* complex; NCBI, National Centre for Biotechnology Information; NS, non-synonymous; OMCL, Ortho Markov Clustering of orthologs; PanGP, pan-genome profile analyze tool; PE, proline-glutamate; PPE, proline-proline glutamate; RAST, rapid annotation using subsystem technology; RD, regions of difference; SNP, single nucleotide polymorphism; SRA, sequence read archive; TB, tuberculosis; UFBoot, ultrafast bootstrap approximation; USA, United States of America; USDA, United States Department of Agriculture; WGS, whole genome sequencing.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Five supplementary tables and six supplementary figures are available with the online version of this article.

that infects the broadest range of host species. It causes animal TB primarily in cattle, but also in other livestock species [3]. Some wildlife species are pathogen reservoirs and infection amplifiers, helping to sustain TB across a variety of ecosystems in different geographic locations [4]. Well-known epidemiological scenarios referred to in the literature include the Iberian Peninsula, United Kingdom, Australia, New Zealand, South Africa and USA [5–7]. Furthermore, *M. bovis* has been associated with zoonotic infection in humans, especially in low-income countries and where the boundaries between humans and animals fade away [8].

The MTC members present a high level of similarity at the nucleotide level, with recent comparative genomics works evidencing an average nucleotide identity >99% [9, 10], however they exhibit variable host tropisms, phenotypes and degrees of pathogenicity [11]. Over the years, genomic diversity analyses have been fundamental to the understanding of evolutionary processes involving MTC as a whole, starting with the study of specific genomic regions, such as the Regions of Difference (RD), proceeding to the phylogeny of specific members, leading to the identification of different *M. tuberculosis* lineages and *M. bovis* clonal complexes [12–16]. Current knowledge supports that *M. bovis* is subdivided into five main clonal complexes: European 1 (Eu1), European 2 (Eu2), European 3 (Eu3), African 1 (Af1) and African 2 (Af2) [12–16]. The clonal complexes are identified by the absence of specific spacers in their spoligotyping profile, specific deletions and single nucleotide polymorphisms (SNPs) in precise genes. Their geographic distribution exhibits differences, with Af1 being predominant in West-Central Africa, Afr2 in East Africa, Eu2 in Iberian Peninsula and Brazil, Eu3 in France and Italy, while Eu1 evidence a worldwide distribution [16–18]. Nevertheless, genome composition and functional diversity, as well as recognition of core- and accessory-gene components, are still overlooked topics for the ecotypes infecting animals.

The pan-genome consists of the entire gene repertoire of a given species, and its analysis is a very informative approach, enabling stratification into the core-genome, containing genes shared by all strains, and the accessory-genome, covering dispensable genes that are not present in all strains, as well as unique genes, also known as singletons, that are particular to specific strains [19]. The core-genome determines the backbone of all strains of a given species and contains genes necessary to the most basic intraspecific biological processes, being responsible for the major phenotype [19]. The accessory-genome adds genetic diversity to the species and may involve other metabolic pathways that crucially assure adaptation to different ecological niches, colonization of new hosts or other functions that are beneficial over other species [19, 20].

The differentiation of genetic variants, grounded by multi-genome studies enabled by whole-genome sequencing, has become central in the field of disease epidemiology, contributing to gaining insights in host–pathogen interactions, virulence, transmission and resistance determinants [21, 22]. In MTC, the combined effect of deletion/insertion events together with the presence of SNPs across members contribute to differences in host ranges, pathogenicity, disease progression and outcome. In fact, previous works performing comparative genome, proteome or secretome analyses of *M. tuberculosis* H37Rv, *M. bovis* and *M. bovis* variant BCG (bacillus Calmette-Guérin) revealed important differences among these ecotypes [20, 23, 24]. For instance, work by Pelayo and collaborators identified about 700 SNPs that could differentiate *M. bovis* BCG and *M. bovis* strains [25].

Proteins or genes related to virulence are involved in a diversity of pathways and biological processes, including the interaction with host cells and interference with immune responses; survival inside the aggressive micro-environment of host macrophages; infection recalcitrance and recrudescence; response to environmental stress conditions; and auxiliary antibiotic resistance [26]. The genotypic and phenotypic differences revealed by the comparative works performed so far encourage further studies on *M. bovis* genomes, which may be insightful regarding the biological and virulence traits of a pathogen that is far from being understood or controlled. The identification of virulence-associated genes, their biological function and the relationship between virulence-specific and epidemiological-specific signatures are important to understand TB infection, persistence and improve control.

In this study, we aimed to gain insights into the *M. bovis* genome structure, composition and evolution by means of a large comparative study at the genome level. For this purpose, 70 *M. bovis* representing the entire clonal complex diversity were submitted to a comparative genome analysis

Impact Statement

Mycobacterium bovis, the causative agent of animal tuberculosis (TB), is proposed to evolve as a conservative strictly clonal microorganism. In the current work, a dataset composed of 70 *M. bovis* representing the entire clonal diversity was used to assess genome structure, composition and evolution. Plus, in a strain subset (*n*=42) circulating in an endemic multi-host TB scenario, the polymorphisms in virulence-related genes were scrutinized. The obtained results support an open and dynamic pan-genome architecture, with a small core-genome component when compared to the accessory-genome component. The diversity of the accessory-genome component might be of crucial importance for *M. bovis* adaptive capacity. The complementary ancestral reconstruction evidenced an overall gene loss supplanting gene gain, with the majority of gene gains occurring in terminal branches. Globally, we provide insights into the evolutionary history of *M. bovis* on a population level and clarify clade-specific virulence signatures of field isolates. Contrary to the paradigmatically view on conservative MTC genomes, this work evidences striking genomic diversity among *M. bovis* field strains.

through a pan-genomic approach, evaluating the genome composition dynamics and performing an ancestral reconstruction of gene gain and loss. Moreover, in a group of *M. bovis* field isolates circulating in an endemic multi-host TB scenario, the polymorphisms in virulence-related genes were ascertained and mutational rates inferred. Globally, we show that molecular approaches to the pan-genome are helpful to gain insights into the evolutionary history of *M. bovis* on a population level and to clarify clade-specific virulence signatures of *M. bovis* field isolates. We then combine this information with phenotypic advantage inferences that may shed light on the evolutionary forces exerted upon this pathogen at the interface with the host in particular epidemiological scenarios.

METHODOLOGY

M. bovis dataset

A globally diverse *M. bovis* dataset was collected and used in this study. Sixteen genomes were downloaded as complete/draft genome assemblies up to a maximum of ten scaffolds from NCBI (National Centre for Biotechnology Information); 12 were selected as representatives of the entire *M. bovis* clonal complex diversity [27] and recovered as Illumina fastq files deposited at SRA (Sequence Read Archive); 42 *M. bovis* genomes were newly sequenced genomes (Table 1 and Table S1, available in the online version of the article, details below). *M. bovis* BCG (bacillus Calmette-Guérin) was excluded from the NCBI search and *M. bovis* AF2122/97, used as reference genome, was included in this dataset. A detailed account of the dataset, including accession code, sequencing strategy, status of completeness and assembly parameters is presented in Table S1.

Altogether, the entire dataset includes isolates recovered from eight host species distributed by 12 countries, collected between 1985 and 2016 and classified into the clonal complexes previously described [16, 17].

Newly sequenced genomes (dataset from Portugal)

The 42 newly sequenced *M. bovis* were entirely recovered from a well-characterized animal TB hotspot in Portugal and included isolates recovered from cattle (*n*=14), red deer (*n*=16) and wild boar (*n*=12) over a 12 year period (circa 70% were recovered over a 5 year period) (Table 1 and Table S1) [28]. The selected host species are described as the most relevant to the maintenance of *M. bovis* in a multi-host system in Portugal [7].

Cultures used in this work were derived from frozen stocks, prepared after a single *in vitro* passage of original archived samples. Re-culture and DNA extraction procedures were performed at a biosecurity level three facility. The frozen culture stocks of selected *M. bovis* were regrown in Middlebrook 7H9 (Difco) medium supplemented with 5% sodium pyruvate and 10% ADS enrichment (50 g albumin, 20 g glucose, 8.5 g sodium chloride in 1 l water), at 37 °C, being monitored regularly for growth. Culture pellet was resuspended in 500 µl PBS, inactivated by heating at 99 °C during 30 min, and then stored at −20 °C until WGS.

WGS was performed using the Illumina Genome Analyser, according to the manufacturer's specifications with the paired-end module attachment. The MiSeq (2×250 pb) technology was implemented for 40 samples at United States Department of Agriculture (USDA, USA), while HiSeq (2×150 pb) technology was employed for the remaining two (Eurofins Genomics, Germany).

Information concerning the microscopic characterization of TB-compatible lesions was available for 39 infected animals (Table 1). Lesions were histologically classified into three categories according with disease progression: type I – 'caseous granulomatous lesions'; type II – 'caseous granulomatous lesions with calcification'; and type III – 'thick encapsulated caseous granulomatous lesions with calcification'.

Bioinformatics workflow

The bioinformatics workflow followed in this work is detailed in supplementary material (Fig. S1) and comprises *de novo* assembly and map to reference strategies.

Genome assembly and annotation

Fifty-four genomes (42 newly sequenced and 12 fastq files recovered from SRA) were *de novo* assembled with Unicycler pipeline, currently available at https://github.com/rrwick/Unicycler [29]. After the reads' quality analysis (FastQC version 0.11.7; https://github.com/s-andrews/FastQC), and whenever necessary after clean-up with Trimmomatic version 0.36 (http://www.usadellab.org/cms/?page=trimmomatic), genomes were assembled with SPAdes optimizer [29], and subjected to post-assembly optimization with Pilon version 1.18 [30]. A conservative bridging mode was applied, since it avoids misassembly, and the *k*-mer size was searched and selected between 20–95% of read length. Considering the reads' size, contigs of less than 300 bp were removed, and a 20-fold cut-off was stabilized following SPAdes guidelines [31].

The QUAST pipeline (http://quast.sourceforge.net/quast.html), which promotes the remapping of contigs with *M. bovis* AF2122/97 reference genome (NCBI accession number LT708304.1), was used to address the quality of the *de novo* assemblies (Table S1).

All the complete genomes/draft assemblies were annotated with the Rapid Annotation using Subsystem Technology (RAST) toolkit, available at webserver https://rast.nmpdr.org [32] using the taxon 1765 ('*Mycobacterium bovis*') and the genetic code 11.

Genome mapping and SNP variant calling

The fastq files of the dataset from Portugal were further submitted to a map to sequence approach. Reads were aligned to the *M. bovis* AF2122/97 reference genome (LT708304.1), using BWA and Samtools [33, 34] through the vSNP pipeline, currently available at https://github.com/USDA-VS/vSNP. Base quality-score recalibration, SNP and

Table 1. Characteristics of *M. bovis* genomes used in this work

M. bovis ID	Clonal complex*	Country	Year	Host species	SNP clade†	Lesion type	Reference	Type of sequence
Mb0220	w/o CC	Portugal	2003	Cattle	E	NA	[65]	Newly sequenced
Mb0261	Eu2	Portugal	2006	Red deer	D	Type I	[65]	Newly sequenced
Mb0601	Eu2	Portugal	2007	Cattle	B	NA	[65]	Newly sequenced
Mb0769	Eu2	Portugal	2008	Cattle	A	Type II	[65]	Newly sequenced
Mb0783	Eu2	Portugal	2008	Wild boar	A	Type II	[65]	Newly sequenced
Mb0865	Eu2	Portugal	2008	Cattle	C	Type I	[65]	Newly sequenced
Mb0891	Eu2	Portugal	2009	Red deer	A	Type III	[65]	Newly sequenced
Mb0893	Eu2	Portugal	2008	Wild boar	B	Type II	[65]	Newly sequenced
Mb1317	Eu2	Portugal	2010	Cattle	D	Type II	[65]	Newly sequenced
Mb1339	Eu2	Portugal	2010	Cattle	A	Type II	[65]	Newly sequenced
Mb1458	w/o CC	Portugal	2010	Wild boar	E	Type II	[65]	Newly sequenced
Mb1480	w/o CC	Portugal	2010	Cattle	E	Type II	[65]	Newly sequenced
Mb1654	Eu2	Portugal	2011	Cattle	A	Type II	[65]	Newly sequenced
Mb1670	w/o CC	Portugal	2011	Red deer	E	Type III	[65]	Newly sequenced
Mb1711	Eu2	Portugal	2011	Red deer	A	Type III	[65]	Newly sequenced
Mb1712	Eu2	Portugal	2011	Red deer	C	Type III	[65]	Newly sequenced
Mb1714	Eu2	Portugal	2011	Cattle	D	Type II	[65]	Newly sequenced
Mb1744	w/o CC	Portugal	2012	Wild boar	E	Type II	[65]	Newly sequenced
Mb1746	Eu2	Portugal	2012	Red deer	B	Type I	[65]	Newly sequenced
Mb1758	Eu2	Portugal	2012	Cattle	A	Type II	[65]	Newly sequenced
Mb1769	Eu2	Portugal	2012	Wild boar	C	Type II	[65]	Newly sequenced
Mb1785	Eu2	Portugal	2012	Red deer	B	Type II	[65]	Newly sequenced
Mb1789	Eu2	Portugal	2012	Cattle	A	NA	[65]	Newly sequenced
Mb1841	Eu2	Portugal	2012	Cattle	A	Type II	[65]	Newly sequenced
Mb1870	Eu2	Portugal	2012	Wild boar	A	Type II	[65]	Newly sequenced
Mb1915	Eu2	Portugal	2013	Red deer	D	Type III	[65]	Newly sequenced
Mb1948	w/o CC	Portugal	2013	Red deer	E	Type III	[65]	Newly sequenced
Mb1960	Eu2	Portugal	2013	Red deer	A	Type I	[65]	Newly sequenced

Continued

Table 1. Continued

M. bovis ID	Clonal complex*	Country	Year	Host species	SNP clade†	Lesion type	Reference	Type of sequence
Mb2026	Eu2	Portugal	2013	Cattle	B	Type II	[65]	Newly sequenced
Mb2043	Eu2	Portugal	2013	Red deer	A	Type III	[65]	Newly sequenced
Mb2067	Eu2	Portugal	2013	Wild boar	B	Type II	[65]	Newly sequenced
Mb2206	Eu2	Portugal	2014	Cattle	B	Type III	[65]	Newly sequenced
Mb2235	w/o CC	Portugal	2014	Red deer	E	Type III	[65]	Newly sequenced
Mb2277	w/o CC	Portugal	2014	Red deer	E	Type II	[65]	Newly sequenced
Mb2300	Eu2	Portugal	2014	Wild boar	B	Type II	[65]	Newly sequenced
Mb2310	Eu2	Portugal	2015	Red deer	A	Type I	[65]	Newly sequenced
Mb2313	Eu2	Portugal	2015	Wild boar	D	Type II	[65]	Newly sequenced
Mb2325	Eu2	Portugal	2015	Red deer	D	Type III	[65]	Newly sequenced
Mb2328	Eu2	Portugal	2015	Red deer	A	Type III	[65]	Newly sequenced
Mb2347	w/o CC	Portugal	2015	Wild boar	E	Type II	[65]	Newly sequenced
Mb2395	Eu2	Portugal	2015	Wild boar	B	Type II	[65]	Newly sequenced
Mb2397	Eu2	Portugal	2015	Wild boar	B	Type III	[65]	Newly sequenced
Mb502499	Af1	Ghana	NA	Human	NA	NA	[27, 66]	SRA deposited
Mb502526	Af1	Ghana	NA	Human	NA	NA	[27, 66]	SRA deposited
Mb1203064	Af1	Ghana	NA	Human	NA	NA	[27, 66]	SRA deposited
Mb4117155	Af2	France	NA	Wild boar	NA	NA	[27, 67]	SRA deposited
Mb1791710	Af2	Tanzania	NA	Chimpanzee	NA	NA	[27, 68]	SRA deposited
Mb1791712	Af2	Tanzania	NA	Chimpanzee	NA	NA	[27, 68]	SRA deposited
Mb1792006	Eu1	USA	2006	Cattle	NA	NA	[68]	SRA deposited
Mb1792127	Eu1	USA	2008	Cattle	NA	NA	[68]	SRA deposited
Mb1792361	Eu1	USA	2013	Cattle	NA	NA	[68]	SRA deposited
Mb7240242	Eu1	USA	2016	Cattle	NA	NA	[68]	SRA deposited
Mb7240415	Eu1	USA	2014	Cattle	NA	NA	[68]	SRA deposited
Mb1791984	Eu1	USA	2005	Cattle	NA	NA	[68]	SRA deposited
MBE1	w/o CC	Egypt	2014	Cattle	NA	NA	NA	assemble/draft genomes NCBI
MBE3	w/o CC	Egypt	2014	Cattle	NA	NA	NA	assemble/draft genomes NCBI

Continued

Table 1. Continued

M. bovis ID	Clonal complex*	Country	Year	Host species	SNP clade†	Lesion type	Reference	Type of sequence
MBE4	w/o CC	Egypt	2014	Cattle	NA	NA	NA	assemble/draft genomes NCBI
MBE10	w/o CC	Egypt	2015	Cattle	NA	NA	NA	assemble/draft genomes NCBI
Mb0077	w/o CC	Canada	2006	Elk	NA	NA	NA	assemble/draft genomes NCBI
Mb0565	w/o CC	Canada	2011	Cattle	NA	NA	NA	assemble/draft genomes NCBI
BMR25	w/o CC	Canada	1985	Bison	NA	NA	NA	assemble/draft genomes NCBI
Mb3601	Eu3	France	2014	Cattle	NA	NA	[16]	assemble/draft genomes NCBI
Mb0476	Eu2	Canada	2002	Cattle	NA	NA	NA	assemble/draft genomes NCBI
MbSP38	Eu2	Brazil	2010	Cattle	NA	NA	[69]	assemble/draft genomes NCBI
Mb1595	w/o CC	Korea	2012	Cattle	NA	NA	[70]	assemble/draft genomes NCBI
Mb0030	w/o CC	China	NA	NA	NA	NA	[71]	assemble/draft genomes NCBI
Mb0001	Eu2	Brazil	2015	Tapirus terrestris	NA	NA	NA	assemble/draft genomes NCBI
Mb0003	w/o CC	India	1986	Cattle	NA	NA	NA	assemble/draft genomes NCBI
Mb31150	Af2	Uganda	NA	Chimpanzee	NA	NA	[27, 72]	assemble/draft genomes NCBI
MbAF2122/97	Eu1	UK	1997	Cattle	NA	NA	NA	assemble/draft genomes NCBI

*Eu1: European 1, Eu2: European 2, Eu3: European 3, Af1: African 1, Af2: African 2, and w/o CC: without clonal complex.
†SNP clade according with [66].
NA, non-available information;

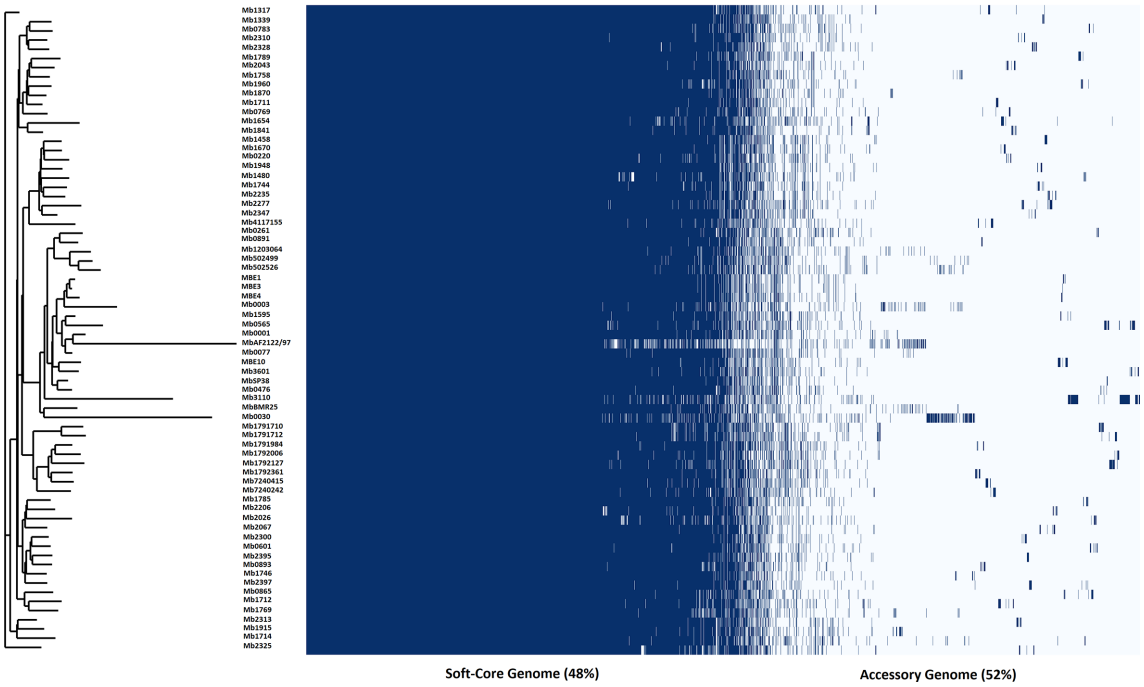

Fig. 1. Soft-core and accessory-genome composition of 70 *M. bovis*. A maximum-likelihood phylogeny built based on pan-genome presence/absence matrix is represented on the left, and a heat map showing gene presence (dark blue) or absence (light blue) is on the right.

indel (insertion or deletion) discovery were applied across all isolates using standard filtering parameters or variant quality-score recalibration according to Genome Analysis Toolkit (GATK)'s Best Practices recommendations [35, 36]. Results were filtered using a minimum SAMtools quality score of 150 and AC=2.

A variant was filtered out if: (i) it was supported by less than 20 reads, (ii) it was found in a frequency of less than 0.9, (iii) it was registered in at least one strain but also with a gap in at least another strain, in order to avoid mapping errors and false SNPs.

Integrated Genomics Viewer (IGV) version 2.4.19 (http://software.broadinstitute.org/software/igv/) was employed to visually validate SNPs and positions with mapping issues or alignment problems. SNPs that fell within Proline-Glutamate (PE) and Proline-Proline Glutamate (PPE) genes were filtered from the analysis, as well as indels.

The sequencing statistics details are provided in Table S2.

The phylogenetic analysis was conducted with MEGA (Molecular Evolutionary Genetics Analysis) software version 7.0 (https://www.megasoftware.net/) using a concatenated sequence of 1816 bp validated and polymorphic SNPs by applying the maximum-likelihood method with 1000 bootstrap inferences and the GTR (General Time Reversible) model.

Pan-genome analyses

Get-homologues pipeline (https://github.com/eead-csic-compbio/get_homologues) [37] was used to compute core- and pan-genome from the input of complete genomes/draft *de novo* assemblies genome sequences.

Briefly, the source GenBank-formatted files were passed to get_homologues.pl script and instructed to compute nucleotide and protein clusters, by running the Clusters of Orthologous Groups triangles (COGtriangles) and OrthoMCL (Markov Clustering of orthologs, OMCL) algorithms, as previously detailed [38]. PFAM-domain scanning was enabled (-D flag).

The directories holding the results from the different clustering algorithms were then passed to the auxiliary script compare_clusters.pl to compute either the consensus core-genome or pan-genome.

Since a large number of draft genomes was used in this work, a 75% of sequence alignment coverage and 95% of sequence identity was imposed and all clusters independent of size were included. Moreover, genes were only considered as taxon specific (clusters of size 1) if sequences had no blast hit with any other protein (E-value 1.0E-05) [39]. This strategy was implemented to minimize the inclusion of truncated genes found at the ends of contigs of open genomes. Truncated genes found at the ends of contigs of incomplete genomes are thus not

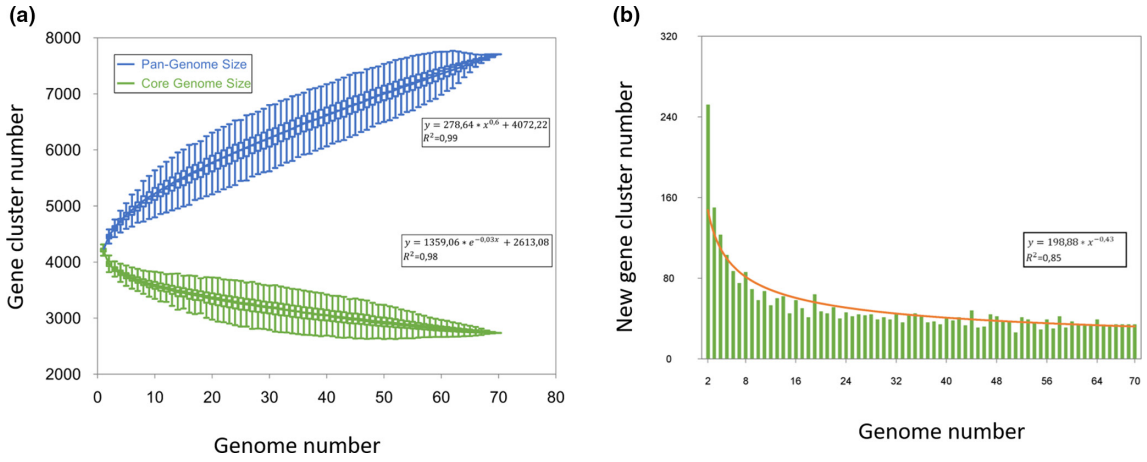

Fig. 2. Prediction of *M. bovis* pan-genome size. (a) Genome size evolution curves of the pan-genome (blue) and core-genome (green). The blue boxes denote the *M. bovis* pan-genome size for each genome for comparison, whereas the green boxes show the same comparison for core-genome size. The curve is the least-squares fit of the power law for the average values. (b) Curve (orange) for the number of new genes with an increase in the number of *M. bovis* genomes.

included in any cluster because they have very weak or no reciprocal BLAST hit.

The auxiliary script parse_pangenome_matrix.pl was employed to perform a pan-genome matrix and to classify genes in four categories: core, soft-core, cloud and shell.

The pan- and core-genome curves and the new genes were depicted using the Pan-Genome Profile Analyse Tool (PanGP) software, using distance-guide (DG) algorithm and default parameters. According with Tettelin's review on pan-genome research [40], Heaps' law model was employed to fit the pan-genome size of strains. The auxiliary script roary_plots.py of Roary pipeline (https://github.com/sanger-pathogens/Roary/blob/master/README.md) [41] was implemented to generate a graphical output, using the pan-genome phylogenetic tree and pan-genome matrix as input files. The scoary pipeline (https://github.com/AdmiralenOla/Scoary) [42] was implemented to infer gene enrichment. The pan-genome matrix, pan-genome phylogenetic tree and a trait file grouping genomes by clonal complex were used as input files. Default criteria were applied to run the pipeline script. The Benjamini–Hochberg adjusted *P*-value (*P*-value<0.05) was used to validate the genes most over- and under-represented in the specific groups.

Maximum-likelihood pan-genome phylogenetic tree from the pan-genome matrix

The auxiliary script, estimate_pangenome_phylogenies.sh, included in get-phylomarkers pipeline (https://github.com/vinuesa/get_phylomarkers) [43] was used to perform a customized maximum-likelihood (ML) phylogenetic analysis, based on the pan-genome matrix files, returned by compare_clusters.pl script (options '-t 0 –m') from the get-homologues pipeline. Maximum-likelihood phylogenetic tree was constructed by IQ-tree with ultrafast bootstrap approximation (UFBoot) of 1000 replicates. The GTR (General Time Reversible)+FO (optimized base frequencies by maximum likelihood)+R3 (unconstrained base frequencies) model was pointed as the most suitable evolutionary model.

Genome evolution – gene gain and loss dynamics

To investigate genome evolution and dynamics across the genealogy of *M. bovis,* a ML birth-and-death model was implemented with the software Count version 10.04 (http://www.iro.umontreal.ca/~csuros/gene_content/count.html) [44], using as input files the ML pan-genome phylogeny and the pan-genome matrix of gene presence/absence.

Briefly, first the model was optimized by maximizing the likelihood of the data using a gain–loss-duplication model with a Poisson distribution for gene family size at the root, and assuming a gamma-distributed rate variation across gene families, and a variable gain/loss ratio across branches. The remaining parameters were settled as default. The rate parameters were optimized after 100 rounds with a convergence threshold of 0.01.

Then, after the optimization of the branch-specific parameters of the model, the ancestral reconstruction was performed using the dollo parsimony method, and gain and loss events were inferred. For each branch i, the gene turnover Ti was defined as Ti= Gi/Li with Gi denoting the branch rate of gene gains and Li, the rate of gene losses.

Functional annotation analysis

For functional annotation analysis, COG categories were associated to the predicted genome CDS via eggnog-mapper v2 webtool (http://eggnog-mapper.embl.de/) [45]. Each CDS was classified into one or more of the 26 upper COG categories. An auto- adjust of the taxonomic scope was enabled and

an e-value cut-off of 1e-3 was settled for orthologue detection. The other parameters were settled as default.

Polymorphic analyses of virulence-related genes

To scrutinize *M. bovis* virulence traits, 421 genes described to be essential for virulence of MTC members, as reviewed by Forrellad and collaborators [26] and/or listed in Mycobrowser database (https://mycobrowser.epfl.ch/), were screened for the presence of SNPs (Table S3). Moreover, due to their biological function in signal transduction, gene regulation and fitness, all genes described to be part of the *mce* family, Lux-R family or to codify toxin and antitoxin proteins were included (Table S3).

Globally, the genes with a demonstrated, or putative role, in mycobacteria virulence were grouped into 13 categories according to their function and molecular features: lipid and fatty acid metabolism (*n*=38), cell-envelope proteins (*n*=14), *mce* family (*n*=33), lipoproteins (*n*=6), secretion systems (*n*=54), defence mechanisms (*n*=33), protein kinases (*n*=3), proteases (*n*=7), metal-transported proteins (*n*=7), genes and metabolism regulation expression (*n*=23), Lux-R family (*n*=6), toxin-antitoxin family (*n*=120), and other virulent proteins (*n*=77) (Table S3).

The SNPs were analysed according to two criteria: (1) number of identified polymorphisms per gene and (2) monomorphic polymorphisms specific to *M. bovis* phylogenetic clades.

Phylogenetic analyses were performed with concatenated alignments of SNPs in the whole set of virulence genes (*n*=195) and concatenated alignments of non-synonymous (NS) SNPs harboured by virulence genes (*n*=119). The maximum-likelihood trees were conducted in MEGA (v 7.0) with 1000 bootstrap inferences and GTR model.

RESULTS AND DISCUSSION

Pan-genome analysis shows an *M. bovis* open pan-genome

Bacterial pan-genome estimation requires *de novo* assemblies into complete or partial genomes with large contigs, providing an unprecedented opportunity to reveal differences when comparing with reference-guided genome assemblies. A total of 54 *M. bovis* isolates were *de novo* assembled using the same sequencing principles and techniques, while the remaining 16 were downloaded from NCBI as complete genome or draft assemblies (Table S1). The seven draft assemblies derived from the public domain were sequenced using the same sequencing technology and assembled through the implementation of two analytical methodologies (Table S1). Excluding the complete genomes (*n*=9), an average of 113 contigs per draft genome was obtained (Table S1).

The 70 complete genome/draft assemblies ranged between 4.17 and 4.36 Mbp, revealing similar genome sizes, and presenting an average GC content of 65.5% (Table S1). The variation in genome size might be the foundation of morphological and physiological differences between strains, contributing to genomic diversity. The predicted number of protein-coding sequences (CDS) pool ranged from 4199 to 4435 (Table S1). The CDS were further clustered in homologue gene clusters and used to estimate the *M. bovis* core- and pan-genomes.

The *M. bovis* genome components were defined as core genome (i.e. genes present in all genomes), soft-core genome [i.e. genes present in at least 95% of the genomes (*n*=66)], cloud genome (i.e. rare genes present only in one/two genomes) and shell genome (i.e. the remaining genes present in several genomes). Shell and cloud formed the accessory component of the genome that contributes to species diversity and may confer selective advantages, such as niche adaptation, antibiotic resistance or colonization of a new host [46]. In this *M. bovis* dataset, the core-genome is composed of 2736 gene clusters, the soft-core by 3708, the shell genome by 1341 and the cloud genome by 2656 (Figs 1 and S2). The accessory component includes 3997 gene clusters, 2656 of which are present only in one or two genomes. On average, each genome has 42 cloud gene clusters, accounting for an average of less than 1% of the total genome-coding capacity. The bimodal, asymmetrical U-shape distribution of gene clusters that is common to many bacterial species indicates that most genes are either rare or to be found in almost all genomes, leading to a smaller proportion of genes at intermediate frequencies (Fig. 2). Our results are in agreement with previous works performed with *M. bovis*, *M. tuberculosis* and combined datasets that evidenced that the accessory component comprise a higher percentage of pan-genome, from circa 21% in a reported dataset with 13 *M. bovis* to 75% in a combined *M. bovis*/*M. tuberculosis* dataset [47–49]. Moreover, studies performed with mycobacterial genomes have also shown that mycobacteria have a small core-genome component when compared to the accessory-genome component [50, 51].

The relative proportions of softcore and accessory components per genome is relatively constant, with each *M. bovis* genome presenting on average (mean and standard deviation) 87.5±0.9% of soft-core genes, 11.5±0.73% of shell genes, and 1.0±1.12% of cloud genes (Fig. S3). Our results are in line with the work of Yang and collaborators, which analysed 13 *M. bovis* strains in which circa 91% of the genome was occupied by core genes [47]; and with the work of Bolotin and Hershberg that associated lower frequencies (circa 2–6%) of cloud genes to bacterial clonal species, including MTC [52].

Core- and accessory-genome size evolution

Core- and pan-genome evolution size estimates were analysed for exponential decay and with growth models (Fig. 2a). According with the obtained growth model, *M. bovis* seems to have an open pan-genome, since there is no distinctly sharp plateau, suggesting that pan-genome size could continue to increase if the number of genomes would also increase (Fig. 2a). Moreover, on average, 32 genes were added by each genome (Fig. 2b). These results are in agreement with the few published works on this topic regarding *M. tuberculosis*

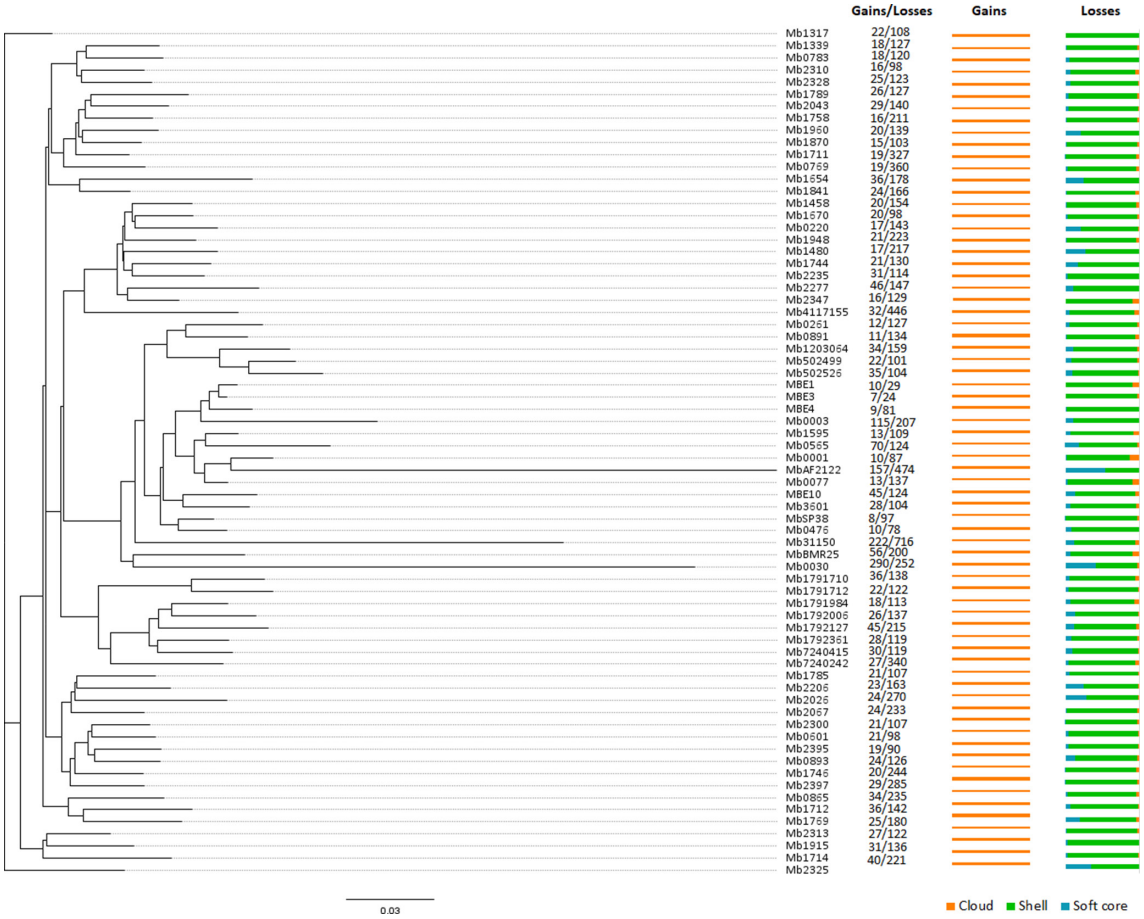

Fig. 3. *M. bovis* genome composition dynamics. Estimated events of genome content evolution are mapped onto the ML phylogeny (GTR) estimated with the pan-genome matrix of gene presence/absence. Numbers, after *M. bovis* labelling, at terminal nodes represent the estimated numbers of gene gain and loss events (Gains/Losses). The two horizontal bars are histograms that represent the relative proportions of soft core (blue), shell (green) and cloud (orange) genes involved in gain (left) and loss (right) events inferred for the terminal branches. The tree is rooted and drawn to scale with branch lengths measured as the number of substitutions per site.

[47, 48] and *M. bovis* [47], that also pointed to an open pan-genome. Open pan-genomes are associated with species that colonize multiple environments and that have multiple ways of exchanging genetic material [46]. *M. bovis* is known, among MTC members, to have the widest host range, with livestock and wildlife reservoirs across different ecosystems in different countries. Plus, the capacity of escaping the host immune system is documented, as well as survival in several *in vitro* and *in vivo* environmental conditions. This high biological flexibility together with the evidence for an open genome suggests that the accessory-genome of *M. bovis* could play an essential role in the adaptive responses triggered by this pathogen.

Dynamics of genome composition

To reconstruct the ancestral dynamics of genome composition across *M. bovis* genealogy, a birth-and-death likelihood model was applied. The analysis revealed a higher loss rate (overall mean 0.011) when comparing with gain rate (overall mean 0.005), with terminal branches evidencing higher values. The gene turnover criteria exhibit a value inferior to one, with the exception of two terminal branches (Mb1595 and Mb0077) and two internal nodes, where the rates of gene gain and loss were estimated to be equal.

The ancestral reconstruction evidenced an overall gene loss (overall mean 147) over gene gain (overall mean 25), with the majority of gene gains occurring in terminal branches (Fig. 3). The terminal branches presented higher mean values for both events, when comparing with internal nodes (mean gene gain/loss of 34/169 and 15/123, for terminal and internal nodes, respectively). Globally, longer terminal branches show more events, as is evident in the branches leading to Mb31150 or Mb0030 (Fig. 3). Branches leading to genomes included in the clonal complexes Eu1 and Af2 reveal high mean values of gene gain and gene loss, when comparing with the genomes of clonal complexes Eu2 and Af1. However, an expanded dataset with more representatives would be necessary to confirm this trend.

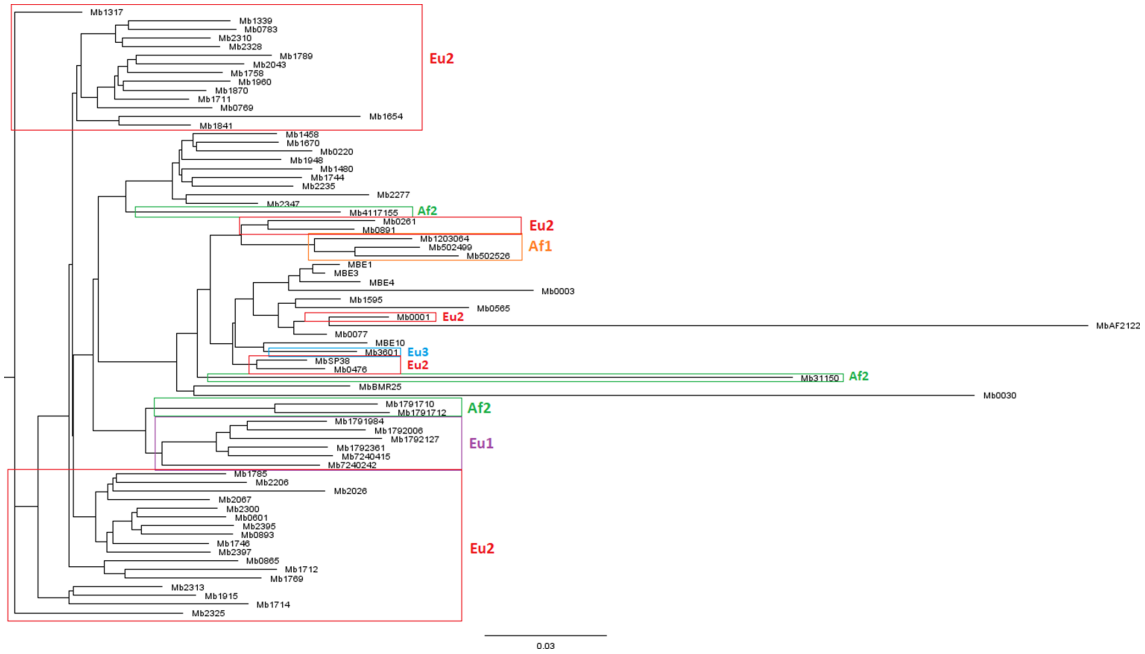

Fig. 4. Maximum-likelihood tree (GTR) of *M. bovis* based on a pan-genome matrix of gene presence/absence. The coloured squares represent the *M. bovis* clonal complexes: purple for European 1 (Eu1), red for European 2 (Eu2), blue for European 3 (Eu3), orange for African 1 (Af1) and green for African 2 (Af2). The tree is rooted and drawn to scale with branch lengths measured as the number of substitutions per site.

The shell component is the most variable, being involved in both gene gain, in internal nodes, and loss events in both terminal and internal nodes, while gains of cloud genes and loss of soft core occur mainly at terminal nodes. We highlight that the loss of strain-specific genes at terminal branches cannot be detected, therefore the loss of cloud genes is underestimated with this methodology. The variance within the cloud component might be of crucial importance for *M. bovis* adaptive capacity. Discrete *M. bovis* subpopulations can thus be perceived based on both core- and accessory-genome components. *M. bovis* exhibits a puzzling discrepancy with high genome nucleotide identity, but high divergence in genome composition. The aforementioned results indicate that (i) diversity within genome composition arises mostly through gene loss, and (ii) recent events that are not shared by many strains need to be considered when investigating *M. bovis* diversity and evolution.

The observed distributions of gain/loss events can either reflect the genealogical process, implying a minimal number of transition steps, or can result from an intricate series of gene gains and losses that likely reflect local selective pressures, such as a tipping point in the ecological niche.

Pan-genome functional annotation

The functions of the predicted CDS encoded in the genomes were inferred from the protein sequences annotated in the COG database. On average, the majority of CDS with COG (85.4%) were attributed to at least one category, while, in a minority of cases, two to five categories could be assigned ('Several categories' group) (Table S1 and Fig. S4). The most represented categories were 'transcription' (total average=258), 'lipid metabolism and transport' (n=242), 'energy production and conversion' (n=214) and 'unknown function' (n=876) (Fig. S4). The variability within the composition of each COG category was estimated to be between zero ('RNA processing and modification') and 24 ('Energy production and conversion'), when comparing the composition of all *M. bovis*.

A gene-enrichment analysis was implemented with the objective of identifying over- and under-represented genes when grouping genomes by clonal complex. After statistical support correction, a higher number of genes was identified in genomes of clonal complex Eu2 (number of genes=171), followed by genomes of Eu1 (n=97), Af1 (n=18) and finally Af2 (n=10). The genes suggested to be over- and under-represented in these clonal complexes were functionally annotated and the categories registering a higher number of occurrences were 'Unknown function' and 'Cell motility' (Fig. S5).

The difference in genome composition among *M. bovis* strains might be driven by selective pressure exerted upon this pathogen, by the environment or the host immune system, however it would be necessary to have tested a larger sample size to further explore this hypothesis.

Pan-genome-based phylogenetic analysis

The phylogenetic analysis based on the presence/absence of pan-genome provided a complementary perspective on the

evolutionary relationships among *M. bovis* strains (Fig. 4). Genomes included in the Af1 (*n*=3) cluster within the same branch, as well as members of Eu1 (*n*=6). These *M. bovis* were recovered from the same epidemiological context, Af1 from human hosts in Ghana and Eu1 from cattle in the USA. Furthermore, two of the four members of Af2 recovered from the same epidemiological scenario cluster together. A higher dataset would thus be necessary to disclose deeper associations.

Impact of the usage of draft genomes on the ability to properly identify CDSs

To further check how the usage of draft genomes might impact the ability to properly identify CDSs, a complementary analysis focused at contig, CDSs and gene gain and loss events was performed. Globally (*n*=70), the genomes with higher CDS prediction were Mb1714 (165 contigs), Mb2313 (119 contigs), Mb0565 (two contigs) and Mb0030 and Mb31150 (both complete genomes). If considering gene-gain events, the higher number of events (over 100 events) was registered by Mb0030, Mb0003, Mb33150 and MbAF2122/8/97 that are complete genomes; while when considering gene loss, the genomes with higher number of inferred events (over 400) were Mb4117155 (99 contigs) and Mb31150 and MbAF2122/97 (both complete genomes) (data not shown).

When considering the group of draft assembly genomes (*n*=61), these presented on average (mean and standard deviation) 113±44 contigs, 4271±11 CDSs, 25±11 gain events and 159±76 loss events. The correlation between the number of contigs per genome and predicted CDSs was weak (*r*=0.29). Moreover, the correlation between the number of contigs per genome and gain and loss events was also weak, with values of 0.04 and 0.21, respectively (data not shown). Our analysis did not point to a positive or negative effect exerted by the number of contigs on the prediction of CDS and gain/loss events, thus supporting that the results described are due to intrinsic strain variability and that our methodological approach to infer core- and pan-genome is robust. The inclusion of truncated genes found at the ends of contigs of open genomes was avoided by adopting a 75% of sequence alignment coverage and 95% of sequence identity criteria, as well as the inclusion of all clusters independently of size, contributed to the robustness of the analyses described herein. Moreover, genes were only considered as taxon specific (clusters of size 1) if sequences had no blast hit with any other protein (E-value 1.0E-05) Finally, when considering SNP identification, a map to reference methodology was implemented, so concerns related to truncated genes did not apply.

Virulence-related genes analyses in *M. bovis* from a multi-host TB system in Portugal

The topology of the maximum-likelihood phylogenetic tree based on the SNP alignment containing 1816 polymorphic positions led to the definition of five *M. bovis* SNP clades (Table 1 and Fig. 5). The phylogenetic tree clearly separates SNP clades A–D (*n*=33), with genomes included in the Eu2 clonal complex, from clade E (*n*=9) that encompasses genomes not assigned to any of the described clonal complexes (Fig. 5). The majority of SNPs (87.1%) was located in coding regions and a prioritized analysis concerning genes involved, or putatively involved, in virulence pathways was performed, leading to the screening of 421 genes (Table S3).

A total of 296 genes were conserved across all analysed strains, while the remaining 125 had at least one polymorphism in one strain, in a total of 194 SNPs. The majority (*n*=118; 60.8%) was non-synonymous (NS) (Figs. 5 and 6a and Table S4). With the exception of the 'lipoproteins' category, all the remaining 12 functional groups comprise polymorphic genes (Fig. 5). Ninety-three genes harbour NS SNPs, 55 had neutral polymorphisms and a group of 23 genes hold both types of SNPs. Considering the NS SNPs, the premature introduction of stop codons in four cases and the loss of a stop codon (Table S5) in one situation could be registered.

Globally, there were SNPs distributed randomly across *M. bovis* strains and SNPs common to strains clustered within the same clade (Table 2, Figs. 5 and 6b). Moreover, an analysis considering host species and geographic location as clustering criteria was performed, but it did not stratify any specific monomorphic SNP positions.

On average, each *M. bovis* strain harboured 35 SNPs in virulence-related genes, with the strains grouping in clade E presenting higher medium values (average SNP number, 53). Each *M. bovis* strain harboured on average 22 NS and 13 synonymous SNPs. The average value of SNPs per virulence-associated gene was greater than one (Fig. S6). The genes *pks12* (*Mb2074c*), with 15 SNPs, *pks5* (*Mb1554c*), with seven SNP, *esxK* (*Mb1229*), with six SNPs, and *esxN* (*Mb1821*) with five SNPs, were the genes that registered the highest number of SNPs (Fig. S6). After normalizing the number of polymorphisms per gene length, the virulence-related genes that exhibit the highest rate of mutation are all included in the *esx* family: *esxK*, *esxN*, *esxM* (*Mb1820*), *esxL* (*Mb1230*) and *esxO* (*Mb2375c*).

Genomic data integration with disease severity was completed using lesion type information. The isolates from SNP clade E are associated to lesion types II and III only, corresponding to more severe disease, while for the remaining SNP clades the lesions types from I to III were registered. Therefore, the SNP clade with the highest number of SNPs in virulence genes is associated with more severe disease outcome. An increase in the whole dataset could contribute to refining this integrative analysis.

Synonymous substitutions are described in the literature as 'silent', once they apparently do not lead to changes in protein sequence. However, some works have suggested that synonymous SNPs can have functional effects, such as decreased mRNA stability and translation [53, 54]. Accordingly, NS polymorphisms are expected to contribute to phenotypic variation due to the putative impacts in protein function, so a meticulous inspection of families with genes harbouring higher rates of mutation, as well as the analysis

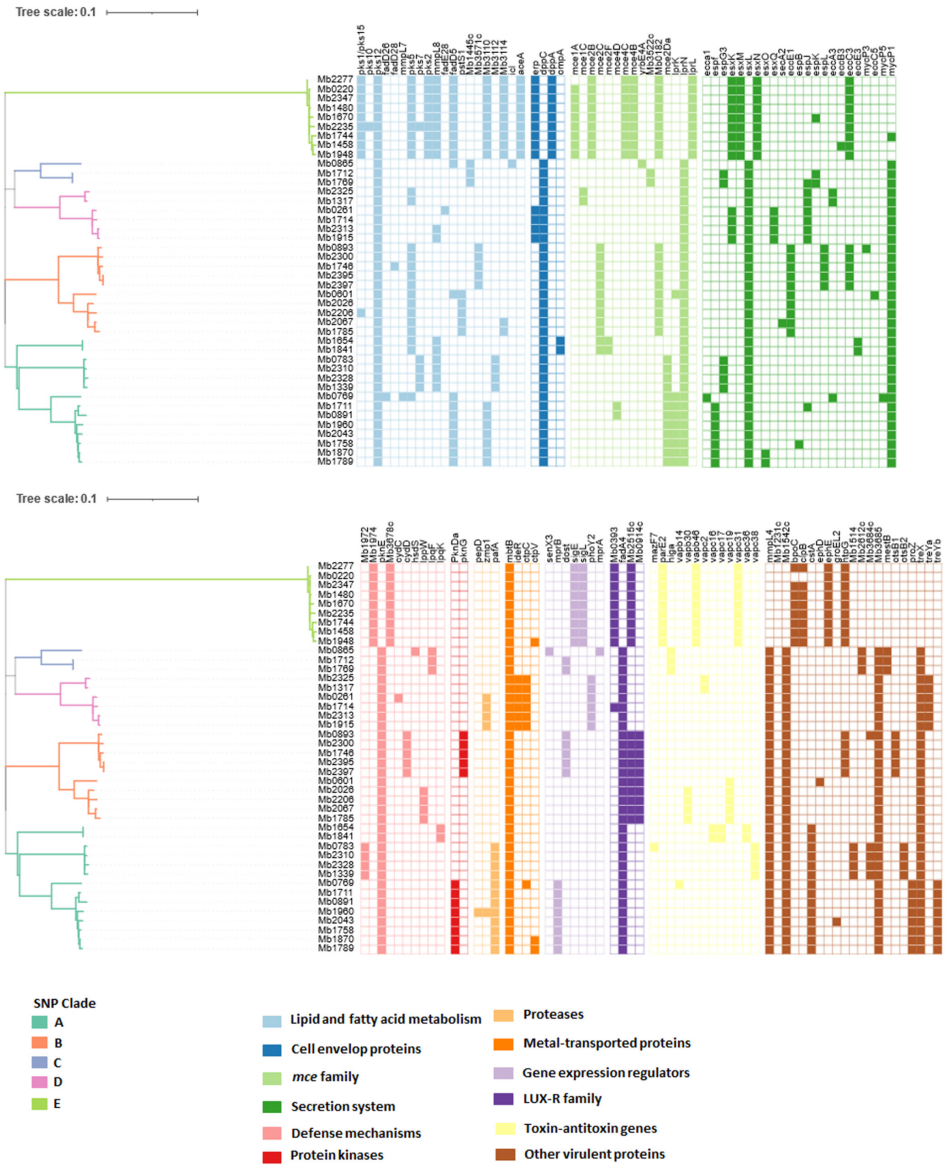

Fig. 5. Maximum-likelihood phylogenetic tree of 42 *M. bovis* strains, from the dataset from Portugal, using as input an alignment containing all validated SNPs (n=1816). The tree is drawn to scale, with branch lengths measured as the number of substitutions per site and identified with different colours according with the SNP clade. The presence/absence SNP profile in virulence-related genes where polymorphisms were recorded (n=125) is represented. Virulence-associated genes are grouped into categories identified by the colour scheme (figure generated with iTOLv6, https://itol.embl.de/).

of clade monomorphic NS SNPs were performed (Table 2 and Fig. 6b). Detailed information regarding gene function and the consequences to mycobacteria of gene deletion are provided in the following sections.

The *esx* gene family exhibits the highest mutation rates

The *esx* genes are members of a specialized secretion system (ESX system) known to enable the transport of antigenic subtracts through the cell wall. This secretion system is associated with pathogenicity and host–pathogen interaction, helping mycobacteria to resist or evade the host immune response [55]. Currently, five ESX systems (ESX-1, ESX-2, ESX-3, ESX-4 and ESX-5) have been described in *M. tuberculosis* and *M. bovis*, with ESX-1, ESX-3 and ESX-5 being required for virulence [55]. The ESX secretion apparatus is constituted by ESX-conserved components (encoded by the *ecc* and *mycP* genes), ESX-type-specific secretion-associated proteins (encoded by the *esp* genes), and secreted/exported proteins (encoded by the *esx* genes and/or PE and PPE proteins) [55, 56]. This *M. bovis* dataset harboured SNPs across the three systems (ESX-1, ESX-3 and ESX-5), as well as in genes included in the three components of the ESX secretion apparatus. Five *esx* genes (*esxK*, *esxM*, *esxL*, *esxN*

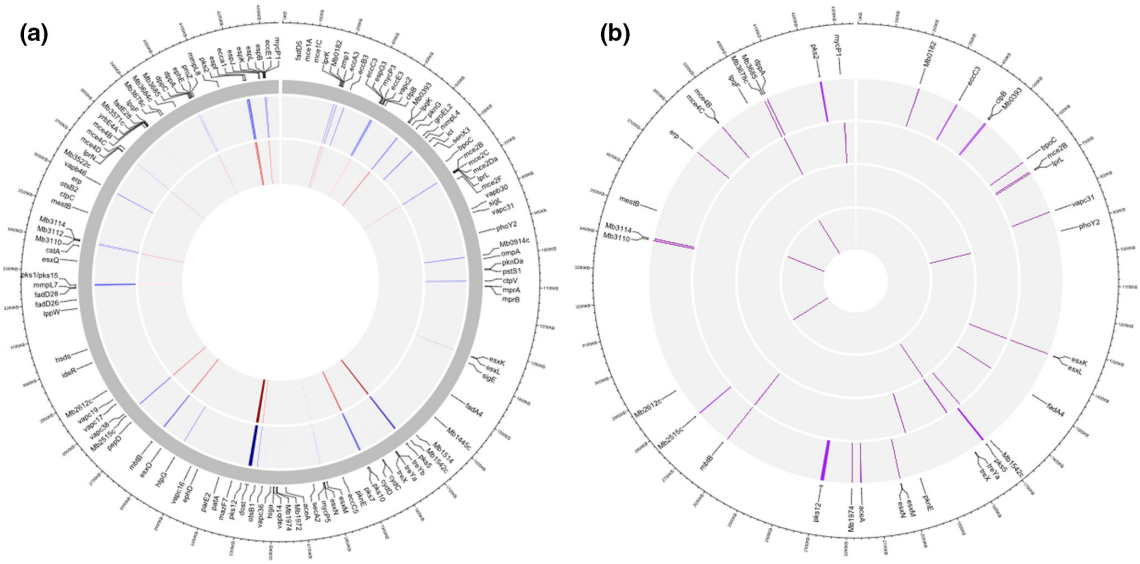

Fig. 6. (a) Circular representation of genome position of genes with SNPs (outer circle) and genes with NS SNPs (inner circle). (b) Circular representation of genome position of genes with monomorphic NS SNPs. From the inner circle to outer circle: clade C, clade D, clade A–D and clade E (figure generated with ShinyCircos, https://venyao.xyz/shinyCircos/).

and *esxO*) had NS SNPs, as did eight conserved components (*eccA1*, *eccE1*, *eccA3*, *eccB3*, *eccD3*, *eccE3* and *mycP1*, *mycP3*), and five secreted associated components (*espG3*, *espB*, *espJ*, *espK*, *espL*) (Table S5). Some of these NS SNPs, in *esx* and conserved component genes were monomorphic to clade E and clades A to D (Table 2).

Polymorphisms in the *esxK*, *esxM* and *esxN* genes, in a total of four positions, are monomorphic for clade E (Table 2 and Table S5). These genes have been proposed to be excreted via the ESX-5 system [55, 56]. Data published so far associate this system to nutrient uptake essential for mycobacteria viability, and to the secretion of PE and PPE proteins of diverse subtypes, with evidences for interference in the pathogenic potential of mycobacteria [55, 57]. Experimental infection works with *M. tuberculosis* deletion mutants for five *esx*-5-encoded PE and PPE proteins showed attenuated virulence in mouse models of animal infection [58]. Further functional works would be needed to explore

Table 2. Identification of SNPs in virulence-associated genes, discriminated by SNP clade

SNP clade	Total SNP sites in virulence genes	Total NS SNP sites	Clade-monomorphic NS SNP sites*	Genes with monomorphic NS SNPs
A (*n*=14)	67	41	–	–
B (*n*=10)	72	45	–	–
C (*n*=3)	34	23	3	*Mb2612c* (*n*=1) *mesTb* (*n*=1) *lpqF* (*n*=1)
D (*n*=6)	37	21	2	*PhoY2* (*n*=1) *treYa* (*n*=1)
A to D (*n*=33)	142	86	8	*esxL* (*n*=1) *pknE* (*n*=1) *mycp1* (*n*=1) *mbtB* (*n*=1) *Mb1542c* (*n*=1) *treX* (*n*=1) *Mb3685* (*n*=1) *fadA4* (*n*=1)
E (*n*=9)	58	35	26	*pks2* (*n*=1) *pks5* (*n*=2) *pks12* (*n*=1) *aceA* (*n*=1), *Mb3110* (*n*=1) *Mb3114* (*n*=1) *erp* (*n*=1) *dppA* (*n*=1) *Mb0182* (*n*=1) *mce4C* (*n*=1) *mce4B* (*n*=1) *mce2B* (*n*=1) *lprL* (*n*=1) *esxK* (*n*=1) *esxM* (*n*=1) *esxN* (*n*=2) *Mb1974* (*n*=1) *Mb3678c* (*n*=1) *mbtB* (*n*=1) *clpB* (*n*=1) *bpoC* (*n*=1) *Mb0393* (*n*=1) *vapc31* (*n*=1) *Mb2515c* (*n*=1) *eccC3* (*n*=1)

*Polymorphic positions leading to non-synonymous alteration present only in the clade-members and common to all.
NS, non-synonymous.;

if the combined phenotype caused by the polymorphisms in these genes impacts *M. bovis* virulence.

The mutations in *esxL* (*n*=1) and *mycP1* (*n*=1) are monomorphic to the remaining 33 strains (clades A to D), while other polymorphisms are singletons or restricted to two strains (Table S3 and S5). The protein EsxL forms a complex with EsxK and their excretion via ESX-5 might be dependent on that association [56]. A reduced intracellular growth in human macrophages was reported in *M. tuberculosis esxKL* deletion mutant, when comparing with the wild-type strain, indicating the involvement in mycobacteria intracellular growth [59]). Since *esxL* is highly expressed during active infection in human lungs [59], polymorphisms might be prejudicial to the complex formation and consequently interfere with the normal excretion pattern.

The gene *mycP1* is part of ESX-1 complex, and previous reports suggest a dual role in substrate processing and the regulation of ESX-1 secretion [60]. The mutagenesis of its active site leads to an increase in secretion, while the *M. tuberculosis* deletion mutant reveals the loss of ESX-1 secretion function [60]. The ESX-1 exported substrates [i.e. proteins ESAT-6, culture filtrate protein 10 (CFP-10) and EspA] are described as major virulence determinants of *M. tuberculosis* during infection, playing a key role in the escape from host defence systems [61].

Clade-specific non-synonymous (NS) monomorphic SNPs

Multiple NS monomorphic SNPs were associated to SNP clades, in a total of three for clade C, two for clade D and 27 for clade E (Table 2 and Table S5). Moreover, if considering clades A to D as a unique group, it was also possible to identify eight monomorphic NS SNPs (Table 2 and STable S5). Detailed information concerning these genes affected by SNPs and non-SNPs is provided in the Supplementary Material (Text S1).

FINAL REMARKS AND FUTURE WORK

The study of genome composition and structure of pathogen populations might provide important insights into their evolutionary dynamics and pathogenic potential. In recent years, many advances concerning *M. bovis* research have been achieved. However, the narrow WGS works available have been mainly applied for transmission inferences in livestock, livestock-wildlife and livestock-human systems. The present work complements previous studies [47, 48]. A dataset composed of 70 strains was used to get insights into the evolutionary forces exerted upon *M. bovis*, with pan-genome and functional classification approaches and detailed analyses concerning gene gain/loss and mutations in genes associated with virulence traits. Not only the genomic composition and diversity of *M. bovis* was detailed but also clade-specific virulence polymorphism signatures were identified in field strains from Portugal.

Contrary to the paradigmatically view on conservative MTC genomes, this work evidences striking genomic diversity among *M. bovis* field strains, providing insights into the core- and accessory-genomes and presenting evidences for an open and dynamic pan-genome. Since *M. bovis* is exposed to dynamic hostile conditions inside the host and also moves across hosts, the accessory component of the genome, especially the cloud moiety that is almost exclusively involved in gene-gain events, could contribute to selective advantages.

The fluctuations in genome composition, with gene gain and loss, together with specific polymorphisms, may act as evolutionary driving forces, facilitating diversification and adaptation. However, the hypothetical advantages conferred by the genome dynamics within some hosts or particular ecosystems (i.e. specific region) needs extra investigation.

We further examined the genetic diversity of virulence-related genes, especially those expected to contribute to phenotypic variation, in the selected group of *M. bovis* from Portugal. Differential pathogenicity abilities might influence the maintenance of strains with specific traits within the ecosystem, due to their capacity to establish a persistent infection in the host or to increased transmissibility. About 30% of the virulence-associated genes that were screened presented a SNP. Plus, it was possible to identify strain-specific NS SNPs, as well as clade monomorphic NS SNPs for clades C, D, E and A–D. Moreover, an association between clade E, that exhibits the higher average number of SNPs in virulence-related genes and also the highest number of monomorphic NS SNPs, along with worse clinical outcome (using lesions of type II and III as proxy for disease severity), needs to be further comprehended by means of a larger dataset. Moreover, a larger and more complete dataset would also allow robust assessment of the relationship between identified SNP-based clades and various metadata, such as host species or geographic location, which would be quite important for understanding the bigger picture of *M. bovis* evolution and adaptation to different hosts. Mutation frequencies varied significantly along nucleotide sequences related with virulence, such that they often concentrated at certain positions. This is the case of the *esx* family that included the genes with the highest mutational rate. Special attention should also be drawn to the *pks* family and *mce* operon that together with the *esx* gene family hold the higher number of NS SNPs for clade E members. The relevance of these polymorphisms to *M. bovis* evolution and TB disease are likely to be discovered in the future as more post-genomic analyses develop.

Dedicated polymorphism analysis of selected genome regions, such as those related with virulence, may inform not only on the intrinsic properties of the mutation process but might also reflect structural and functional features to be explored under several perspectives. This information combined with experimental functional works may link genotype to phenotype, clarifying the consequences of such SNPs and the underlying adaptive advantages for

mycobacteria, potentially opening new avenues for control strategies in the long term. Most functional works focus on *M. tuberculosis*, so it is crucial to investigate the functional consequences of *M. bovis* genotypic diversity associated with virulence factors, particularly some specific genes/gene families, and the subsequent clinical presentation of animal TB. This future research might help to explain the infecting success of certain strains in multi-host systems and might be crucial to redefine control actions potentially opening new possibilities, such as wildlife vaccination. Currently, *M. bovis* BCG remains the single licensed anti-tuberculosis vaccine against human TB. In the animal TB field, the implementation of wildlife vaccination has been explored through experimental trials in badger (*Meles meles*) and wild boar (*Sus scrofa*) in Europe, brushtail possum (*Trichosurus vulpecula*) in New Zealand, or white-tailed deer (*Odocoileus virginianus*) in the USA using *M. bovis* BCG or heat-inactivated *M. bovis* [62–64]. A deeper knowledge of the genetic diversity and peculiarities of *M. bovis* field strains, including evolutionary dynamics, virulence and pathogenicity, is crucial to customize effective *M. bovis* vaccines or other approaches for animal TB control in each ecosystem.

Funding Information
This work was funded by Programa Operacional de Competitividade e Internacionalização (POCI) (FEDER component), Programa Operacional Regional de Lisboa, and Fundação para a Ciência e a Tecnologia (FCT), Portugal, in the scope of project 'Colossus: Control Of tuberCuLOsiS at the wildlife/livestock interface uSing innovative natUre-based Solutions' (ref. POCI-01-0145- FEDER- 029783) and strategic funding to cE3c and BioISI Research Units (UIDB/00329/2020 and UIDB/04046/2020). ACR was supported by FCT through doctoral grant (PD/BD/128031/2016).

Author contributions
M.V.C. conceived this work and secured funding. A.C.R. performed the bioinformatic analyses and explored data under the supervision of M.V.C. A.C.R. wrote the first draft of the manuscript and M.V.C. critically revised and redirected all drafts. Both authors approved the final version.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
The panel of *M. bovis* isolates analysed here was selected for WGS from a wider *M. bovis* dataset recovered in Portugal [64] in the scope of official control plans for animal TB and research projects. No animals were sacrificed for the purposes of this study. None of the authors were responsible for the death of any animals nor were any samples used in the study collected by the authors. All applicable institutional and/or national/international guidelines for the care and use of animals have been followed.

References

1. Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* 2018:202–213.
2. WHO. *Global tuberculosis report*. 2019.
3. Reis AC, Ramos B, Pereira AC, Cunha M. Global trends of epidemiological research in livestock tuberculosis for the last four decades. *Transbound Emerg Dis* 2020;1–14.
4. Reis AC, Ramos B, Pereira AC, Cunha M. The hard numbers of tuberculosis epidemiology in wildlife: A meta-regression and systematic review. *Transbound Emerg Dis* 2020;9:1–20.
5. Corner LAL. The role of wild animal populations in the epidemiology of tuberculosis in domestic animals: How to assess the risk. *Vet Microbiol* 2006;112:303–312.
6. Palmer M, Thacker TC, Waters WR, Gortázar C, Corner LAL. Mycobacterium bovis: A model pathogen at the interface of livestock, wildlife, and humans. *Vet Med Int* 2012;2012:236205.
7. Cunha M, Monteiro M, Carvalho P, Mendonça P, Albuquerque T, *et al*. Multihost tuberculosis: insights from the portuguese control program. *Vet Med Int* 2011;2011:795165.
8. Luciano SA, Roess A. Human zoonotic tuberculosis and livestock exposure in low- and middle-income countries: A systematic review identifying challenges in laboratory diagnosis. *Zoonoses Public Health* 2020;67:97–111.
9. Riojas MA, McGough KJ, Rider-Riojas CJ, Rastogi N, Hazbón MH. Phylogenomic analysis of the species of the *Mycobacterium tuberculosis* complex demonstrates that *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium caprae*, *Mycobacterium microti* and *Mycobacterium pinnipedii* are later heterotypic synonyms of mycob. *Int J Syst Evol Microbiol* 2018;68:324–332.
10. Jia X, Yang L, Dong M, Chen S, Lv L, *et al*. The bioinformatics analysis of comparative genomics of Mycobacterium tuberculosis complex (MTBC) provides insight into dissimilarities between intraspecific groups differing in host association, virulence, and epitope diversity. *Front Cell Infect Microbiol* 2017:7.
11. Brosch R, Gordon S, Marmiesse M, Brodin P, Buchrieser C, *et al*. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci U S A* 2002;99:3684–3689.
12. Smith NH, Berg S, Dale J, Allen A, Rodriguez S, *et al*. European 1: A globally important clonal complex of *Mycobacterium bovis*. *Infect Genet Evol* 2011;11:1340–1351.
13. Rodriguez-Campos S, Schürch AC, Dale J, Lohan AJ, Cunha M, *et al*. European 2 – A clonal complex of Mycobacterium bovis dominant in the Iberian Peninsula. *Infect Genet Evol* 2012;12:866–872.
14. Muller B, Hilty M, Berg S, Garcia-Pelayo MC, Dale J, *et al*. African 1, an Epidemiologically Important Clonal Complex of Mycobacterium bovis Dominant in Mali, Nigeria, Cameroon, and Chad. *J Bacteriol* 2009;191:1951–1960.
15. Berg S, Garcia-Pelayo M, Muller B, Hailu E, Asiimwe B, *et al*. African 2, a Clonal Complex of *Mycobacterium bovis* Epidemiologically Important in East Africa. *J Bacteriol* 2011;193:670–678.
16. Branger M, Loux V, Cochard T, Boschiroli ML, Biet F, *et al*. The complete genome sequence of Mycobacterium bovis Mb3601, a SB0120 spoligotype strain representative of a new clonal group. *Infect Genet Evol* 2020;82:104309.
17. Rodriguez-Campos S, Smith NH, Boniotti MB, Aranaz A. Overview and phylogeny of Mycobacterium tuberculosis complex organisms: Implications for diagnostics and legislation of bovine tuberculosis. *Res Vet Sci* 2014;97:S5-19.
18. É S, de Alencar A, Hodon M, Filho S, de Souza-Filho A, *et al*. Identification of clonal complexes of Mycobacterium bovis in Brazil. *Arch Microbiol* 2019;201:1047–1051.
19. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol* 2015;23:148–154.
20. Golby P, Nunez J, Witney A, Hinds J, Quail MA, *et al*. Genome-level analyses of *Mycobacterium bovis* lineages reveal the role of SNPs and antisense transcription in differential gene expression. *BMC Genomics* 2013;14:710.
21. Cheng G, Hussain T, Sabir N, Ni J, Li M, *et al*. Comparative study of the molecular basis of pathogenicity of M. Bovis strains in a mouse model. *Int J Mol Sci* 2019.
22. Dong H, Lv Y, Sreevatsan S, Zhao D, Zhou X. Differences in pathogenicity of three animal isolates of *Mycobacterium* species in a mouse model. *PLoS One* 2017;12:e0183666.
23. Vargas-Romero F, Mendoza-Hernández G, Suárez-Güemes F, Hernández-Pando R, Castañón-Arreola M. Secretome profiling of highly virulent mycobacterium bovis 04-303 strain reveals higher abundance of virulence-associated proteins. *Microbial Pathogenesis* 2016;100:305–311.

24. Mattow J, Schaible UE, Schmidt F, Hagens K, Siejak F, *et al*. Comparative proteome analysis of culture supernatant proteins from virulent *Mycobacterium tuberculosis* H37Rv and attenuated M.*bovis* BCG Copenhagen. *Electrophoresis* 2003;24:3405–3420.
25. Pelayo MCG, Uplekar S, Keniry A, Lopez PM, Garnier T, *et al*. A Comprehensive Survey of Single Nucleotide Polymorphisms (SNPs) across *Mycobacterium bovis* Strains and *M. bovis* BCG Vaccine Strains Refines the Genealogy and Defines a Minimal Set of SNPs That Separate Virulent *M. bovis* Strains and *M. bovis* BCG Strains. *Infect Immun* 2009;77:2230–2238.
26. Forrellad MA, Klepp LI, Gioffré A, Sabio y García J, Morbidoni HR, *et al*. Virulence factors of the Mycobacterium tuberculosis complex. *Virulence* 2013;4:3–66.
27. Zimpel CK, Patané JSL, Guedes ACP, de Souza RF, Silva-pereira TT, *et al*. Global Distribution and Evolution of *Mycobacterium bovis* Lineages. *Front Microbiol* 2020;11:843.
28. Reis AC, Tenreiro R, Albuquerque T, Botelho A, Cunha M. Long-term molecular surveillance provides clues on a cattle origin for mycobacterium bovis in Portugal. *Sci Rep* 2020;10:1–18.
29. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
30. Walker B, Abeel T, Shea T, Priest M, Abouelliel A, *et al*. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* 2014;9:e112963.
31. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al*. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
32. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, *et al*. RASTtk: A modular and extensible implementation of the RAST algorithm for annotating batches of genomes. *Sci Rep* 2015;5:8365.
33. Li H, Durbin R. Fast and accurate short read alignment with Burrows – Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al*. The Sequence Alignment / Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
35. Depristo M, Banks E, Poplin R, Garimella K, Maguire J, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–498.
36. Van der Auwera G, Carneiro MO, Hartl C, Poplin R, del Angel G, *et al*. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma* 2014;43.
37. Contreras-moreira B, Vinuesa P. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Appl Environ Microbiol* 2013;79:7696–7701.
38. Vinuesa P, Contreras-moreira B. Robust Identification of Orthologues and Paralogues for Microbial Pan-Genomics Using GET_HOMOLOGUES: A Case Study of pIncA/C Plasmids. In: *Bacterial Pangenomics*. 2015. pp. 203–232. https://doi.org/10.1007/978-1-4939-1720-4
39. Lefébure T, Stanhope MJ. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biol* 2007;8:71.
40. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008;11:472–477.
41. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
42. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 2016;17:1–9.
43. Vinuesa P, Ochoa-Sánchez LE, Contreras-Moreira B. Get_phylomarkers, a software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus Stenotrophomonas. *Front Microbiol* 2018;9:771.
44. Csűös M. Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 2010;26:1910–1912.
45. Tatusov RL, Galperin MY, Natale DA, Koonin E. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000;28:33–36.
46. Medini D, Donati C, Tettelin H, Mausignani V, Rappuoli R. The microbial pan-genome. *Curr Opin Genet Dev* 2005;15:589–594.
47. Yang T, Zhong J, Zhang J, Li C, Yu X, *et al*. Pan-Genomic Study of Mycobacterium tuberculosis Reflecting the Primary/ Secondary Genes, Generality/ Individuality, and the Interconversion Through Copy Number Variations. *Front Microbiol* 2018;9:1886.
48. Periwal V, Patowary A, Vellarikkal SK, Gupta A, Singh M, *et al*. Comparative Whole-Genome Analysis of Clinical Isolates Reveals Characteristic Architecture of Mycobacterium tuberculosis Pangenome. *PLoS One* 2015;10:e0122979.
49. Zakham F, Sironen T, Vapalahti O, Kant R. Pan and Core Genome Analysis of 183 *Mycobacterium tuberculosis* Strains Revealed a High Inter-Species Diversity among the Human Adapted Strains. *Antibiotics (Basel)* 2021;10:500.
50. Zakham F, Aouane O, Ussery D, Benjouad A, Ennaji M. Computational genomics-proteomics and Phylogeny analysis of twenty one mycobacterial genomes (Tuberculosis & non Tuberculosis strains. *Microb Inform Exp* 2012;2:7.
51. Trost B, Haakensen M, Pittet V, Ziola B, Kusalik A. Analysis and comparison of the pan-genomic properties of sixteen well-characterized bacterial genera. *BMC Microbiol* 2010;10:1–18.
52. Bolotin E, Hershberg R. Gene loss dominates as a source of genetic variation within clonal pathogenic bacterial species. *Genome Biol Evol* 2015;7:2173–2187.
53. Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, *et al*. (n.d.) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum Mol Genet*;2003:205–2016.
54. Capon F, Allen MH, Ameen M, Burden AD, Tillman D, *et al*. A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Hum Mol Genetics* 2004;13:2361–2368.
55. Gröschel MI, Sayes F, Simeone R, Majlessi L, Brosch R. ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat Rev Microbiol* 2016;14:677–691.
56. Bitter W, Houben ENG, Bottai D, Brodin P, Brown EJ, *et al*. Systematic genetic nomenclature for type VII secretion systems. *PLoS Pathog* 2013;5:e1000507.
57. Ates LS, Ummels R, Commandeur S, van der Weerd R, Sparrius M, *et al*. Essential Role of the ESX-5 Secretion System in Outer Membrane Permeability of Pathogenic Mycobacteria. *PLoS Genet* 2015;11:e1005190.
58. Bottai D, di Luca M, Majlessi L, Frigui W, Simeone R, *et al*. Disruption of the ESX-5 system of *Mycobacterium tuberculosis* causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. *Mol Microbiol* 2012;83:1195–1209.
59. Bukka A, Price CTD, Kernodle DS, Graham JE. *Mycobacterium tuberculosis* RNA expression patterns in sputum bacteria indicate secreted Esx factors contributing to growth are highly expressed in active disease. *Front Microbiol* 2012;2:266.
60. Ohol YM, Goetz DH, Chan K, Shiloh MU, Craik CS, *et al*. Mycobacterium tuberculosis MycP1 Protease Plays a Dual Role in Regulation of ESX-1 Secretion and Virulence. *Cell Host Microbe* 2010;7:210–220.
61. Conrad WH, Osman MM, Shanahan JK, Chu F, Takaki KK, *et al*. Mycobacterial ESX-1 secretion system mediates host cell lysis through bacterium contact-dependent gross membrane disruptions. *Proc Natl Acad Sci U S A* 2017;114:1371–1376.
62. Ballesteros C, Garrido JM, Vicente J, Romero B, Galindo RC, *et al*. First data on eurasian wild boar response to oral immunization with BCG and challenge with a mycobacterium bovis field strain. *Vaccine* 2009;27:6662–6668.

63. Nugent G, Yockney IJ, Whitford EJ, Cross ML, Aldwell E, *et al*. Field Trial of an Aerially-Distributed Tuberculosis Vaccine in a Low-Density Wildlife Population of Brushtail Possums (Trichosurus vulpecula. *PLoS One* 2016;11:e0167144.
64. Gormley E, Bhuachalla DN, Keeffe JO, Murphy D, Aldwell FE, *et al*. Oral Vaccination of Free-Living Badgers (Meles meles) with Bacille Calmette Guérin (BCG) Vaccine Confers Protection against Tuberculosis. *PLoS One* 2017;12:e0168851.
65. Reis AC, Salvador LCM, Robbe-Austerman S, Tenreiro R, Botelho A, *et al*. Phylogenomics sheds light on the population structure of mycobacterium bovis from a multi-host tuberculosis system. *bioRxiv* 2021.
66. Otchere ID, van Tonder AJ, Asante-Poku A, Sánchez-Busó L, Coscollá M, *et al*. Molecular epidemiology and whole genome sequencing analysis of clinical Mycobacterium bovis from Ghana. *PLoS One* 2019;14:e0209395.
67. Branger M, Hauer A, Michelet L, Karoui C, Cochard T, *et al*. Draft Genome Sequence of Mycobacterium bovis Strain D-10-02315 Isolated from Wild Boar. *Genome Announc* 2016;4:e01268:16.:.
68. Orloski K, Robbe-Austerman S, Stuber T, Hench B, Schoenbaum M. Whole genome sequencing of *Mycobacterium bovis* isolated from livestock in the United States, 1989-2018. *Front Vet Sci* 2018;5:253.
69. Guimarães AMS, Zimpel CK, Ikuta CY, do Nascimento NC, dos Santos AP, *et al*. Draft genome sequence of *Mycobacterium bovis* strain SP38, a pathogenic bacterium isolated from a bovine in Brazil. *Genome Announc* 2015;3.
70. Kim N, Jang Y, Kim JK, Ryoo S, Kwon KH, *et al*. Complete genome sequence of Mycobacterium bovis clinical strain 1595, isolated from the laryngopharyngeal lymph node of South Korean cattle. *Genome Announc* 2015;3:e01124:15.:.
71. Zhu L, Zhong J, Jia X, Liu G, Kang Y, *et al*. Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Res* 2016;44:730–743.
72. Wanzala SI, Nakavuma J, Travis DA, Kia P, Ogwang S, *et al*. Draft genome sequences of Mycobacterium bovis BZ 31150 and Mycobacterium bovis B2 7505, pathogenic bacteria isolated from archived captive animal bronchial washes and human sputum samples in Uganda. *Genome Announc* 2015;3:11.