






# Construction and Analysis of the Complete Genome Sequence of Leprosy Agent *Mycobacterium lepromatosis*

 Francisco J. Silva,<sup>a,b</sup>  Diego Santos-Garcia,<sup>c</sup> Xiaofeng Zheng,<sup>d</sup> Li Zhang,<sup>d\*</sup>  Xiang Y. Han<sup>e</sup>

<sup>a</sup>Institute for Integrative Systems Biology (I2SysBio), University of Valencia and CSIC, Paterna, Spain

<sup>b</sup>Genomics and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research, Valencia, Spain

<sup>c</sup>Laboratory of Biometry and Evolutionary Biology UMR CNRS, University of Lyon, Villeurbanne, France

<sup>d</sup>Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

<sup>e</sup>Department of Laboratory Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

**ABSTRACT** Leprosy is caused by *Mycobacterium leprae* and *Mycobacterium lepromatosis*. We report construction and analyses of the complete genome sequence of *M. lepromatosis* FJ924. The genome contained 3,271,694 nucleotides to encode 1,789 functional genes and 1,564 pseudogenes. It shared 1,420 genes and 885 pseudogenes (71.4%) with *M. leprae* but differed in 1,281 genes and pseudogenes (28.6%). In phylogeny, the leprosy bacilli started from a most recent common ancestor (MRCA) that diverged ~30 million years ago (Mya) from environmental organism *Mycobacterium haemophilum*. The MRCA then underwent reductive evolution with pseudogenization, gene loss, and chromosomal rearrangements. Analysis of the shared pseudogenes estimated the pseudogenization event ~14 Mya, shortly before species bifurcation. Afterwards, genomic changes occurred to lesser extent in each species. Like *M. leprae*, four major types of highly repetitive sequences were detected in *M. lepromatosis*, contributing to chromosomal rearrangements within and after MRCA. Variations in genes and copy numbers were noted, such as three copies of the gene encoding bifunctional diguanylate cyclase/phosphodiesterase in *M. lepromatosis*, but single copy in *M. leprae*; 6 genes encoding the TetR family transcriptional regulators in *M. lepromatosis*, but 11 such genes in *M. leprae*; presence of *hemW* gene in *M. lepromatosis*, but absence in *M. leprae*; and others. These variations likely aid unique pathogenesis, such as diffuse lepromatous leprosy associated with *M. lepromatosis*, while the shared genomic features should explain the common pathogenesis of dermatitis and neuritis in leprosy. Together, these findings and the genomic data of *M. lepromatosis* may facilitate future research and care for leprosy.

**IMPORTANCE** Leprosy is a dreaded infection that still affects millions of people worldwide. *Mycobacterium lepromatosis* is a recently recognized cause in addition to the well-known *Mycobacterium leprae*. *M. lepromatosis* is likely specific for diffuse lepromatous leprosy, a severe form of the infection and endemic in Mexico. This study constructed and annotated the complete genome sequence of *M. lepromatosis* FJ924 and performed comparative genomic analyses with related mycobacteria. The results afford new and refined insights into the genome size, gene repertoire, pseudogenes, phylogenomic relationship, genome organization and plasticity, process and timing of reductive evolution, and genetic and proteomic basis for pathogenesis. The availability of the complete *M. lepromatosis* genome may prove to be useful for future research and care for the infection.

**KEYWORDS** *Mycobacterium leprae*, *Mycobacterium lepromatosis*, reductive evolution, genomics, leprosy

Leprosy is one of the oldest known human diseases. Despite tremendous advances during the last several decades, this infection continues to be an important health problem in many developing countries. Leprosy affects skin and peripheral nerves chronically and manifests in the clinical forms of tuberculoid, borderline, lepromatous,

**Editor** Gaurav Sharma, Institute of Bioinformatics and Applied Biotechnology

**Copyright** © 2022 Silva et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Francisco J. Silva, francisco.silva@uv.es, or Xiang Y. Han, xhan@mdanderson.org.

\*Present address: Li Zhang, Department of Environmental Health, The University of Cincinnati College of Medicine, Cincinnati, Ohio, USA.

The authors declare no conflict of interest.

**Received** 27 September 2021

**Accepted** 7 April 2022

**Published** 25 April 2022

and diffuse lepromatous leprosy (DLL) (1, 2). Since initial discovery in 1873, *Mycobacterium leprae* has been known to be the sole cause of leprosy (3). In 2008, a new species, named *Mycobacterium lepromatosis*, was found to be a second cause of leprosy in two patients of Mexico origin who died of DLL (4).

Genome sequencing of *M. leprae* has revealed a phenomenon of reductive evolution (5). The genome size (3.27 Mb) was much smaller ( $>1$  Mb) than that of related *Mycobacterium tuberculosis* (6). More surprisingly, it contained around 1,600 pseudogenes with loss of  $\sim 50\%$  of the ancestral genes (7). Comparative analyses of *M. leprae* pseudogenes with their orthologous *M. tuberculosis* genes have led to the proposal that a massive pseudogenization took place  $\sim 20$  million years ago (Mya) (7). The lean *M. leprae* genome has been extraordinary stable, as revealed by genome sequences and typing of many worldwide strains that showed only clonal differences (0.005%) (8).

The clonal differences among *M. leprae* strains contrast a 9.1% genetic difference with *M. lepromatosis* that was revealed through analysis of 20 genes and pseudogenes (22.8 kb) (9). The study also estimated a divergence time of  $\sim 10$  Mya between the two leprosy bacilli. With analysis of the draft genome of *M. lepromatosis* strain Mx1-22A, this divergence was refined to 13.9 Mya (10). The leprosy bacilli are phylogenetically closely related to *Mycobacterium haemophilum* (11), an environmental organism with very low pathogenicity. The genome sequence of *M. haemophilum* (4.24 Mb) shows that almost all coding genes (CDS) are functional in contrast to the decayed genomes of the leprosy bacilli (12).

Clinical and pathological studies on *M. lepromatosis* have been revealing as well. Many studies have identified this agent in patients with leprosy from American and Asian countries, including Mexico (4, 13, 14), Canada (15), Costa Rica (16), Brazil (17), Myanmar (17), and Singapore (18). Some of these studies have also revealed dual agent infections caused by *M. lepromatosis* and *M. leprae* in some patients (14, 16–18). In Mexico, *M. lepromatosis* is likely the dominant cause of leprosy and specific for the endemic DLL (14). DLL is a unique and severe form of lepromatous leprosy occurring in  $\sim 20\%$  of patients (9, 14). It is characterized by diffuse nonnodular cutaneous infiltration and recurrent crops of large and sharply demarcated ischemic skin lesion called Lucio's phenomenon (2, 19). In the advanced stage, lesions may become ulcerated or even generalized, particularly on the lower extremities, leading to fatal secondary infection and sepsis. Pathologically, the mycobacteria invade deep into the subcutis, blood vessels, and internal organs, in addition to skin and nerves – the hallmark of leprosy (14, 20). The vasculitis eventually leads to endothelial proliferation, vascular occlusion, ischemia, and skin necrosis (4, 14). The discovery and further characterization of *M. lepromatosis* should thus enable research and insight into the prominent involvement of subcutis (panniculitis) and blood vessels (leukocytoclastic vasculitis) in DLL.

In Europe where leprosy has long been eliminated, red squirrels on British Isles have been found to be infected with the leprosy bacilli (21). The study further showed that, while the squirrel *M. leprae* strains were most closely related to the medieval human strains, the squirrel *M. lepromatosis* strains had a divergence time of 27,000 years from the Mexican strain Mx1-22A. This divergence time raises an intriguing question as to how the bacterium landed in the British red squirrels.

While the draft genome of *M. lepromatosis* Mx1-22A has characterized the vast majority of genes and pseudogenes (10), the presence of 126 contigs and potential omit of unique component make it desirable to construct and analyze the complete genome sequence of this organism. Here, we report construction and annotation of the complete genome sequence of *M. lepromatosis* strain FJ924 and comparative genomic analyses with related mycobacteria. This study has improved our understanding of the complete evolution of the chromosome of both species, including rearrangements, gene losses and gene duplications. We have also estimated the relative ages of pseudogenes from *M. leprae* and *M. lepromatosis* to time the process of massive pseudogenization in relation to the divergence points.

**TABLE 1** Comparative genomic features of *M. lepromatosis* FJ924 and related mycobacteria<sup>a</sup>

Feature	<i>M. lepromatosis</i> strain		<i>M. leprae</i> strain		<i>M. haemophilum</i>	<i>M. uberis</i> Jura	<i>M. tuberculosis</i>
	FJ924	Mx1-22A	Br4923	TN	DSM 44634		H37Rv
Genome size (bp)	3,271,694	3,206,741	3,268,071	3,268,203	4,235,765	3,122,721	4,411,532
Protein coding genes (CDS)	1,789	1,477	1,604	1,605	3,728	1,759	4,018
CDS pseudogenes	1,564	1,331	1,116	1,115	153	1,081	13
rRNA genes	3	3	3	3	3	2	3
tRNA genes	46	45	45	45	45	44	45
tmRNA gene	1	1	1	1	-	1	1
Other noncoding RNA genes	13	3	3	2	3	3	30
% GC content	57.89	57.89	57.80	57.80	63.90	57.50	65.60
No. contigs	1	126	1	1	1	54	1

<sup>a</sup>Data from GenBank/RefSeq genome. *M. lepromatosis* FJ924, this work.

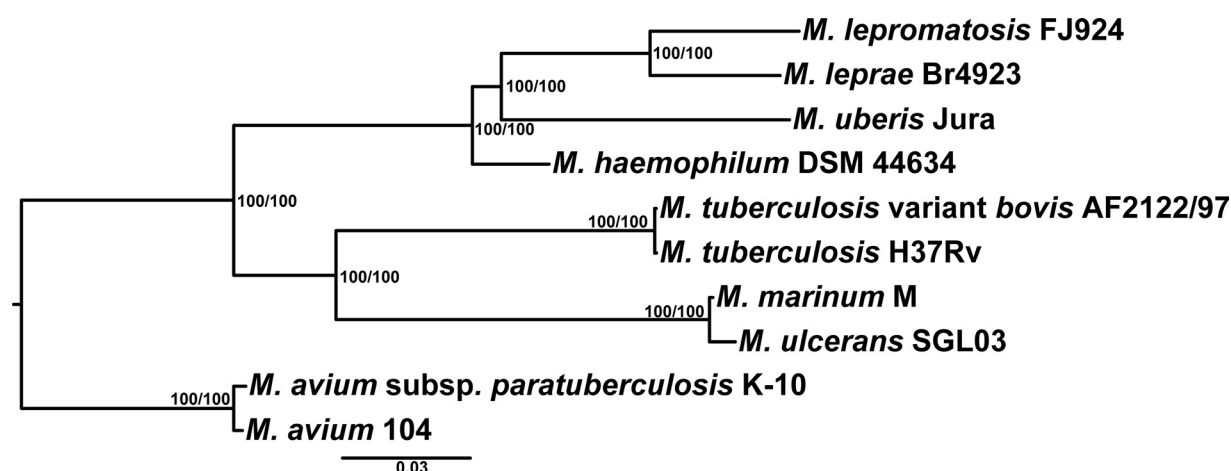
## RESULTS AND DISCUSSION

**Construction of the complete genome.** The construction of *M. lepromatosis* genome encountered several technical hurdles. Due to the lack of axenic cultivation, the quantity and quality of genomic DNA became the first rate-limiting step. The genomic DNA was extracted from a dried, stained, and archived smear slide, with only ~18% of sequenced reads (12.3 of 69 million reads) belonging to *M. lepromatosis*. Contaminant reads included human host DNA and nonspecific bacteria DNA. Some bacterial DNA reads, from *Thermus scotoductus* particularly, were likely introduced and/or amplified during the library preparation step that used polymerase chain reactions (PCR) to boost target quantity. Other nonspecific bacterial DNA, such as from *Propionibacterium acnes*, *Micrococcus luteus*, and others, were likely from the stained smear that used regular reagents for acid-fast stain, tap water, and microscopy immersion oil. To overcome contamination, initial steps used matches to the nearest *M. leprae* genome to capture specific reads, *de novo* assembly, systematic tagging and removal of contaminant contigs, and alignment to *M. leprae* for draft construction. Later more iterative cycles of mapping, *de novo* assembly, and gap filling were used to retrieve unique sequences, close gaps and refine drafts. PCR and Sanger sequencing of amplicons were used to close final gaps, verify assembly of several long contigs, and settle copy numbers of short repetitive sequences. The genome coverage was 525×.

**Genome features and phylogeny.** The seamless genome of *M. lepromatosis* strain FJ924 contained 3,271,694 bp. The genome encoded 1,789 CDS, 1,564 pseudogenes, and 63 noncoding RNA genes (3 rRNA, 46 tRNA, 1 tmRNA and 13 other ncRNA) (Table 1). It was slightly longer, by 3,623 bp (0.1%), than the genome of *M. leprae* Br4923 (3,268,071 bp), but much shorter, by 964,071 bp (29.5%), than the *M. haemophilum* genome (4,235,765 bp). The CDS covered 50.4% of the genome and averaged 922 bp per CDS. The pseudogenes covered 36.4%, averaging 761 bp per pseudogene. Intergenic sequences, including repetitive elements, covered the remaining 13.2%, averaging 129 bp per intergenic spacer.

A phylogenomic tree was constructed based on the proteomes (316,716 alignment sites) of well-known mycobacteria and the newly named and sequenced *Mycobacterium uberis*, a yet-to-be cultivated organism that causes nodular thelitis and tuberculoid scrotoitis in cows and goats (22). As shown in Fig. 1, *M. leprae* and *M. lepromatosis* stem from a most recent common ancestor (MRCA) that forms a clade with *M. haemophilum* and *M. uberis*. Shortly after the divergence of *M. haemophilum*, the lineages of *M. uberis* and the MRCA diverged again and both underwent genome downsizing, pseudogenization, and decrease in GC content (Table 1).

**Comparison with the draft genome of *M. lepromatosis* Mx1-22A.** The FJ924 genome was compared with the draft genome of *M. lepromatosis* Mx1-22A that consisted of 3,206,741 bp from 126 contigs (10). This detected 66 kb extra nucleotides from 56 segments with lengths >300 bp, with the largest segment covering 8,518 bp (Table S1A). Annotation of these segments revealed 27 CDS (21 kb, 1.5% of all CDS) and 65 pseudogenes (44 kb, 4.2% of all pseudogenes). Most of these genes and pseudogenes were absent in Mx1-22A, while a few of them showed partial sequence of the



**FIG 1** Phylogenomic tree of selected mycobacterial species. ML tree was inferred using a concatenated conserved protein alignment (316,716 amino acid positions) under a JTT+F+R3 substitution model. Support values were obtained with 1000 ultrafast bootstraps (right node labels) and 1000 SH-aLRT (left node labels).

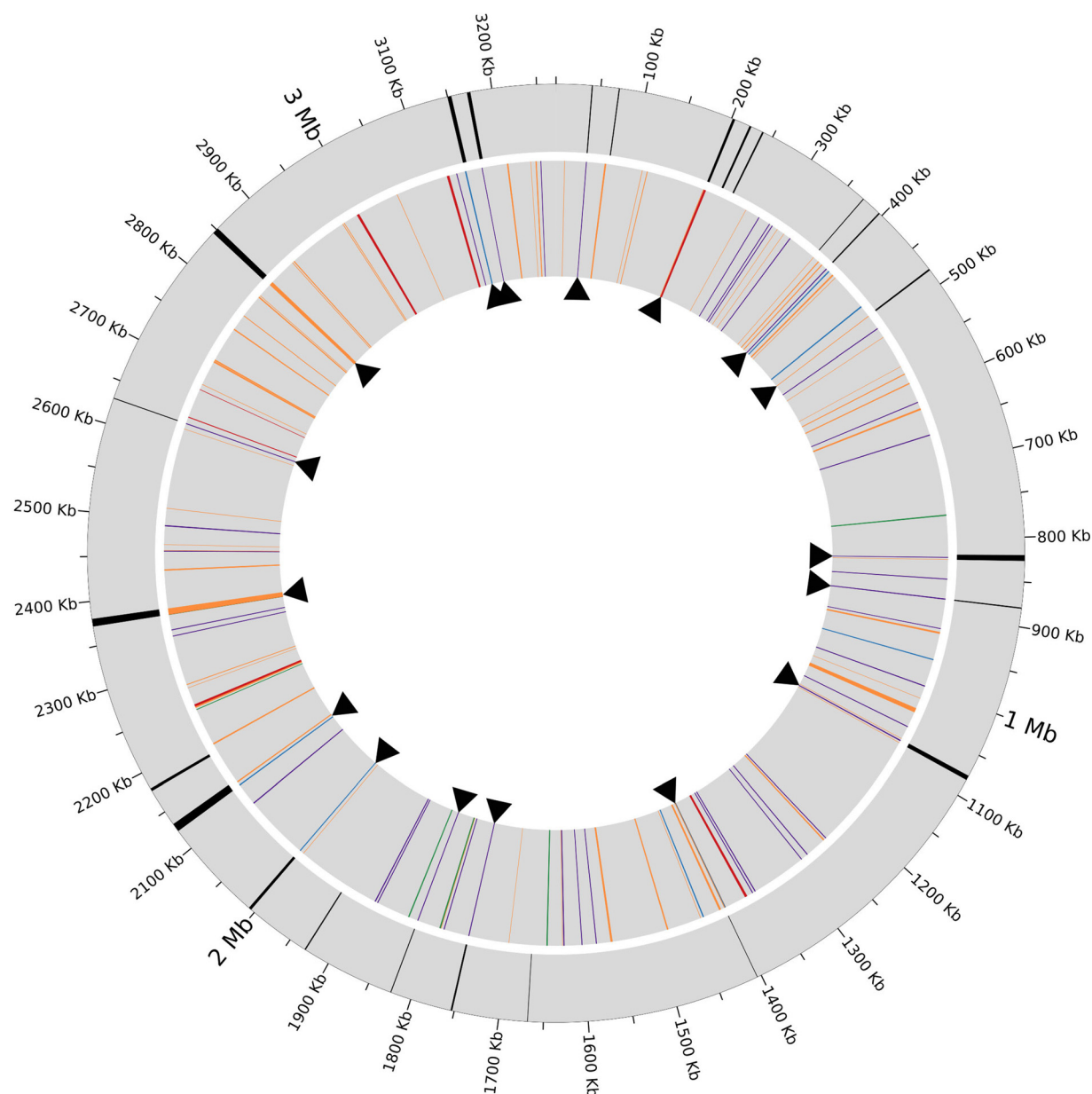
annotated feature (Table S1B and S1C). Shorter Mx1-22A was likely due to loss of some *M. lepromatosis* unique sequences during array capture to the *M. leprae* genome, to the assembler that discards/collapses repetitive regions, and to low sequencing coverage of the genome (55×) (10).

Single nucleotide polymorphisms (SNPs) between Mx1-22A and FJ924 were identified with MAUVE (23, 24). After exclusion of ambiguities and inaccuracies near the edges of Mx1-22A contigs, a final set of 11 SNPs were obtained, with six in genes (five in CDS and one in the 16S rRNA gene) and five in intergenic regions (Table S2). Four SNPs detected in coding genes produced one amino acid difference between the proteins encoded by the two strains. The fifth SNP in a CDS encoded a PPE protein of 410 amino acids in FJ924 (MLPF\_1455), but it caused a stop codon in Mx1-22A, hence a pseudogene (MLPM\_1054). The limited number of SNPs suggests that the divergence between Mx1-22A and FJ924, strains likely from the Monterrey region in north central Mexico and the Sonora region in north western Mexico, respectively (~1000 km away) (4, 10), was very recent, probably a few hundred years ago. Similarly, another *M. lepromatosis* strain PL-02, also from Mexico, diverged from Mx1-22A ~186 years ago (21).

Among the extra CDS detected, the complete FJ924 genome contained two tandemly arranged asparagine permease genes (MLPF\_1849 and 1850), like *M. leprae*, and three identical copies of the gene encoding bifunctional diguanylate cyclase/phosphodiesterase, unlike *M. leprae*, *M. uberis* and *M. tuberculosis* with a single copy (further analysis later). The Mx1-22A contained single asparagine permease gene and single diguanylate cyclase/phosphodiesterase gene (10); such omissions were likely due to relaxed genome assembly and/or lack of copy count instead of true absence. Assembly errors in three Mx1-22A contigs were also seen, such as contig-111, a 14,151-bp fragment, that showed a minus orientation of the first 8,706 bp.

**Repetitive sequences.** The *M. leprae* genome is characterized by the presence of numerous repetitive sequences that likely play a role in genome plasticity (25). Repetitive sequences that appeared twice or more in the *M. lepromatosis* FJ924 genome were analyzed (Table S3). In total, they accounted for 3.5% of the genome and were scattered in 125 genome segments. Some repeats dispersed similarly to the repetitive elements in the *M. leprae* genome (25). Four main families of repetitive elements were identified: types A, B, C and D, with 38, 8, 8 and 6 copies, respectively (Fig. 2 and Table S3A). These repeats were named as RLPM, LPMREP, REPLPM, and LPMRPT to correspond to RLEP, LEPREP, REPLEP, and LEPRPT, respectively, in the *M. leprae* genome.

Among the 38 copies of RLPM of various extension lengths, a core sequence of 626 nucleotides was identified in 18 copies (see consensus in Table S3B). The complete or partial

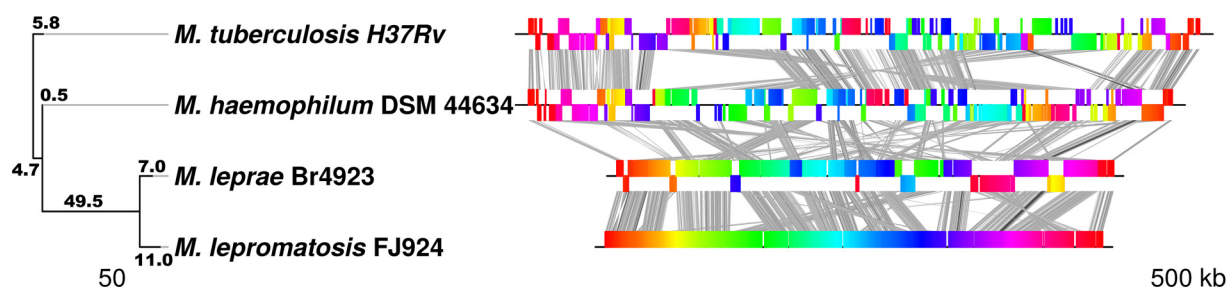


**FIG 2** Synteny breaks and repeat sequences identified in *M. lepromatosis* FJ924. The outside circle displays the 24 synteny breaks (black) detected in the genome comparing with *M. leprae*. The inner circle displays the positions of repeat sequences: RLPM (purple), LPMREP (red), REPLPM (blue), LPMRPT (yellow) repeat families, and other unidentified repeats (orange). Arrowheads mark the coincidences between the positions of breaks and repeat sequences.

core sequences were highly conserved in the 38 copies (99.2–100% identity). However, compared with all 37 RLEP copies in *M. leprae*, RLPM showed low sequence identity, in 25 copies at best 75% of 126 nucleotides with several gap sites. Nonetheless, in view of the nearly identical copy numbers, similar average length (~700 bp), and some homologous genome locations, RLEP and RLPM probably evolved from the same ancestral repeat after divergence from MRCA.

Type B repeats LPMREP resembled the LEPREP repeats in *M. leprae* in BLAST matches (79–83%) and copy numbers. Both showed five complete copies and three fragment copies in their respective genomes. The origin of these repeats seems to be a pseudogene of a putative group II intron maturase-related protein, which persisted from the MRCA. Based on the alignment of the complete copies of LPMREP, a core sequence of 2,453 nucleotides





**FIG 3** Synteny among mycobacterial genomes. (Left) Genome rearrangement phylogeny obtained with the Neighbor joining method and a distance matrix showing the minimal number of inversion events required to explain the differences between any pair of genomes obtained with GRIMM. Branch lengths are the numbers of inversion events estimated with the phylogenetic method. (Right) Graphic linear representation of the genome rearrangements observed comparing the four mycobacterial genomes. The graph was obtained with Mauve and displayed with genoPlotR.

was identified (Table S3B). Recently, these repeats were used in a *M. lepromatosis* molecular diagnostic assay, but under the abbreviation of ‘RLPM’ (16).

Type C repeats REPLPM in *M. lepromatosis* were equivalent to the REPLEP repeats in *M. leprae*, but with gapped alignments and an overall 74% match. The copy numbers also varied: 8 copies for REPLPM but 15 copies for REPLEP (25). The REPLPM copies displayed different lengths but a consensus core of 1,123 nucleotides could be inferred (based on 50% coverage). This core sequence (Table S3B) was completely included in REPLPM\_07.

Type D repeats LPMRPT were equivalent to LEPRPT repeats in *M. leprae* with 87% similarity covering almost the complete sequence. Only four of LPMRPT repeats were large (~1,200 bp) (see LPMRPT\_01 in Table S3B), while the two others were short fragments.

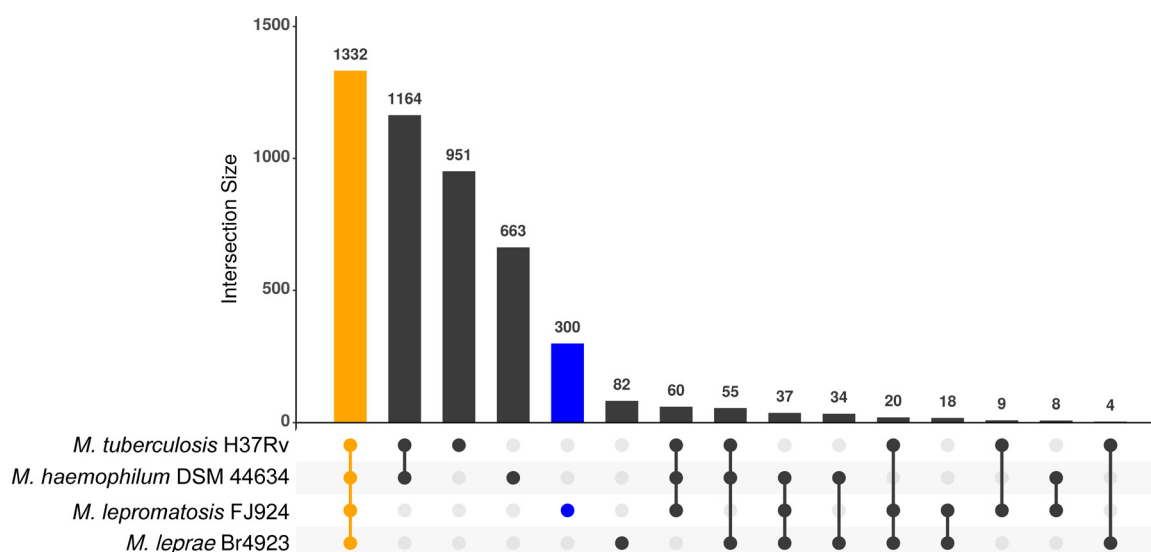
Functionally, the four repetitive elements likely also played a role in genome plasticity and organization (see below). Other repetitive sequences were also annotated (Table S3A). They likely arose from segmental duplications or gene families.

**Comparative genomics: synteny and proteome.** Synteny among the genomes of *M. leprae*, *M. lepromatosis*, *M. haemophilum* and *M. tuberculosis* was compared with the program MAUVE (23) (Fig. 3). Many differences were observed between the two leprosy bacilli, by one way, and between *M. haemophilum* and *M. tuberculosis*, by the other, indicating events of chromosomal rearrangement.

To infer the minimal number of chromosomal rearrangements among them, a pairwise inversion distance matrix was constructed with the program GRIMM (26), which predicted 18 inversion events between *M. leprae* and *M. lepromatosis* and 11 between *M. haemophilum* and *M. tuberculosis* (Fig. 3). In addition, the MRCA of leprosy bacilli incurred ~50 chromosomal rearrangements after divergence from *M. haemophilum*. Comparisons of the consecutive CDS and pseudogenes in *M. leprae* and *M. lepromatosis* further rendered 25 syntenic blocks as the likely outcome of major rearrangement events (Table S4). These blocks and the 24 synteny breaks were analyzed for genesis by comparing them with ancestral *M. haemophilum* and *M. tuberculosis* (Table S5). For example, blocks 4 and 5, and blocks 21 and 22 in *M. lepromatosis* was similarly continuous as in *M. haemophilum* and *M. tuberculosis* whereas they were separated in *M. leprae*. Similarly, by the order of *M. leprae*, seven colinear blocks boundaries aligned with *M. haemophilum* (and with *M. tuberculosis* in six cases) to suggest ancestral origin. Together, these seven *M. leprae* connections affected 11 synteny breaks in *M. lepromatosis* (Table S5).

At the 24 synteny breaks, 17 repetitive elements were noted to be involved: 9 RLPM, 3 REPLPM, 2 LPMREP and 3 other repeats (Fig. 2 and Table S5). These observations reveal the abrupt effect of repeats through recombination to shape the genome of *M. lepromatosis*. These chromosomal rearrangements may also produce gene duplications and deletions and affect gene expression. Similar effects were also previously observed for the genome of *M. leprae* (25).

Orthologous clusters of proteins were estimated with the program OrthoFinder v2.3.3 (27) for proteomes of the four mycobacteria (Fig. 4). A core proteome of 1,332



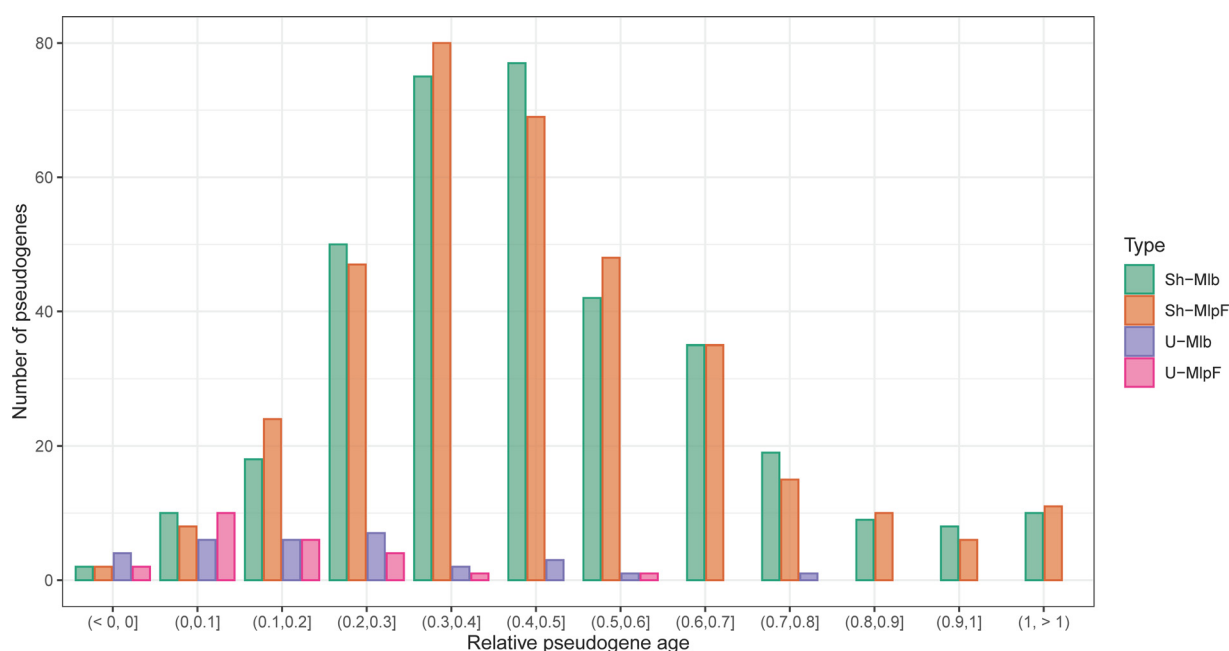
**FIG 4** Comparative analyses of mycobacterial proteomes. UpSet plot showing the computed coding gene clusters (intersections) derived from the proteomes encoded in the complete genomes of *M. lepromatosis* FJ924, *M. leprae* Br4923, *M. haemophilum* DSM44634, and *M. tuberculosis* H37Rv. Shared coding gene clusters (core) and *M. lepromatosis* FJ924 specific coding gene clusters are highlighted in yellow and blue, respectively.

protein clusters, with at least one member encoded in each genome, was detected. Of them, a set of 1,227 clusters included a single protein in each proteome (single copy orthogroups). The number of clusters with only proteins from one leprosy proteome was relatively high (82 in *M. leprae* and 300 in *M. lepromatosis*). However, most of them were annotated as hypothetical proteins, only a few with an annotated function (13 in *M. lepromatosis*). There were also several orthogroups, including proteins from only leprosy proteomes (18 clusters). Among the latter, three showed a functional annotation: MLPF\_0519 (Lamb/YcsF family protein), MLPF\_0609 (PPE family protein) and MLPF\_0662 (lysophospholipase protein domain). In summary, the proteome of *M. lepromatosis* is a reduced version of cultivable mycobacteria with almost no specific functional proteins.

**History of pseudogenization in the leprosy bacilli.** The pseudogenization events in *M. lepromatosis* and *M. leprae*, in terms of the ages of pseudogenes, were analyzed to time the occurrences, i.e., before divergence, after divergence, or both. A set of 355 shared pseudogenes, 24 unique pseudogenes in *M. lepromatosis*, and 30 unique pseudogenes in *M. leprae* were obtained for analysis through annotation and alignment (>200 nucleotide sites) of orthologous functional CDS in *M. haemophilum* and *M. tuberculosis* and CDS or pseudogene in one leprosy bacillus. While the requirement of orthology precluded all pseudogenes for analysis, the qualified ones should render insights into the tempo of pseudogenization.

The relative ages of pseudogenes (*Rpa*) are shown in Fig. 5. Values for unique pseudogenes were small, with mean and SD of  $0.138 \pm 0.139$  and  $0.216 \pm 0.190$  for *M. lepromatosis* and *M. leprae*, respectively, suggesting more recent occurrences after divergence. Values for shared pseudogenes were larger,  $0.462 \pm 0.238$  for *M. lepromatosis* and  $0.467 \pm 0.254$  for *M. leprae* along with a wider and overlapping distribution. An x-y plot of the shared pseudogenes showed a positive correlation ( $R^2 = 0.704$ ) (data not shown), indicating a common pseudogenization process prior to divergence. Together, the analysis of *Rpa* suggests that most of them occurred approximately during the same period as an event of massive pseudogenization.

A time tree analysis using the RelTime method (28) estimated the time of divergence between *M. haemophilum* and the MRCA of the leprosy bacilli to be 29.52 Mya (Fig. 6). When this value was set to be *Rpa* of 1, Mya values for the shared pseudogenes



**FIG 5** Relative ages of pseudogenes. Histogram with the relative ages estimated for four types of pseudogenes. A value of 1 corresponds to the time of divergence between *M. haemophilum* and the leprosy bacilli. A value of 0 is present time. From left to right: *M. leprae* shared pseudogenes (with an orthologous pseudogene in *M. lepromatosis*, green color), *M. lepromatosis* shared pseudogenes (with an orthologous pseudogene in *M. leprae*, orange color), *M. leprae* unique pseudogenes (with a CDS annotated in *M. lepromatosis*, blue color) and *M. lepromatosis* unique pseudogenes (with a CDS annotated in *M. leprae*, pink color).

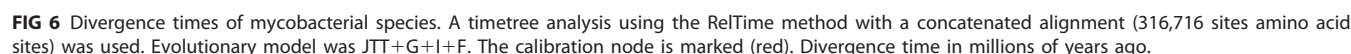
were  $13.64 \pm 7.03$  for *M. lepromatosis* and  $13.79 \pm 7.5$  for *M. leprae*. This result and the positive correlation between the values of shared pseudogenes suggest a massive pseudogenization event occurred close but slightly prior to divergence of the two leprosy bacilli at 13.85 Mya, noted previously (10), but much later than the divergence from the related species *M. ubris* at 28 Mya (Fig. 6).

**Pathogenetic insights from shared genomic features.** The leprosy bacilli have a moderate pathogenicity, between rarely pathogenic *M. haemophilum* and highly pathogenic *M. tuberculosis* (11). The lean genomes and shared genomic features underscore the clinical and pathological features of leprosy, i.e., long incubation period, insidious onset, dermal and neuronal invasion of bacillus-laden macrophages, and chronicity, contrasting to mainly pulmonary infection in tuberculosis.

Table 2 shows the numbers of shared and unique CDS and pseudogenes for *M. leprae* and *M. lepromatosis*. The two bacilli shared 1,420 CDS and 885 pseudogenes (71.4% of genome) but differed in 1,281 CDS and pseudogenes (28.6%), although some of them may be associated with the different annotation protocols. Some shared features were examined to highlight common pathogenesis in leprosy.

Among all known mycobacteria (~160 species), neuronal invasion is unique to the leprosy bacilli. In *M. leprae*, adherence to Schwann cells seems to require the presence of the DNA-binding protein HU and the cell wall antigen phenolic glycolipid 1 (PGL-I) (29). PGL-I also subverts host immune cells to shield the bacillus from clearance (30). The synthesis of PGL-I requires 6 enzymes/genes (31). In *M. lepromatosis* FJ924, the PGL-I genes were identified as MLPF\_0154, 0155, and 0156, and MLPF\_2688, 2689, and 2690 in two clusters without related pseudogenes. The bacillus also contained the gene encoding DNA-binding protein HU (MLPF\_0968) that is also known as mycobacterial DNA binding protein 1 or laminin-binding protein (32). This protein consists of an N-terminal HU-like region and a C-terminal IDR (intrinsically disordered region). The *M. lepromatosis* protein contained 195 residues, slightly smaller than the proteins of *M. leprae*, *M. haemophilum* and *M. tuberculosis* (200, 213 and 214 residues, respectively). The length difference was close to the C terminus at the IDR, essential region for such function as genome compaction or suppression of DNA synthesis (33).





The ESX-5 system plays a pivotal role in the secretion of proteins rich in Pro-Glu (PE proteins) and Pro-Pro-Glu (PPE proteins) (35). *M. lepromatosis* contained mainly the core genes of the system, MLPF\_0876 to 0880 (*eccA*, *eccE*, *mycP*, *eccD* and *espG*) and MLPF\_0889 and 0890 (*eccC* and *eccB*). A pseudogene of the *esat-6* like protein (MLPF\_0881) was detected between these two gene clusters. In the orthologous *M. leprae* region, neither an *esat-6* gene nor a pseudogene was detected. Genes encoding the PE/PPE proteins were examined in the genomes of leprosy bacilli for comparison with those in *M. haemophilum* and *M. tuberculosis*. OrthoFinder clustered these proteins/genes in 10 orthogroups (Table S6). While the *M. haemophilum* and *M. tuberculosis* genomes each contained 46 genes spread in the clusters, *M. lepromatosis* and *M. leprae* each contained only 13 genes but 34 and 32 pseudogenes, respectively (Table S7).

	<i>M. lepromatosis</i> FJ924			
<i>M. leprae</i> Br4923	CDS	Pseudogene	Absent	Total
CDS	1420	128	56	1604
Pseudogene	54	885	177	1116
Absent	315	551		866
Total	1789	1564	233	3586

**TABLE 3** Preferential inactivation of PE/PPE genes but retention of PGL-I genes and ribosomal protein genes in *M. lepromatosis* FJ924

Feature	PE/PPE comparison			PGL-I comparison			Ribosomal proteins		
	PE/PPE	All others	Total	PGL-I	All others	Total	RBS	All others	Total
No. genes	13	1776	1789	6	1783	1789	52	1737	1789
No. pseudogenes	34	1530	1564	0	1564	1564	1	1563	1564
Total	47	3306	3353	6	3347	3353	53	3300	3353
Inactivation odds ratio	3.04			0			0.02		
Statistical test	$\chi^2 = 12.6, P = 0.0004$			Fisher's $P = 0.033$			$\chi^2 = 43.3, P < 0.0001$		

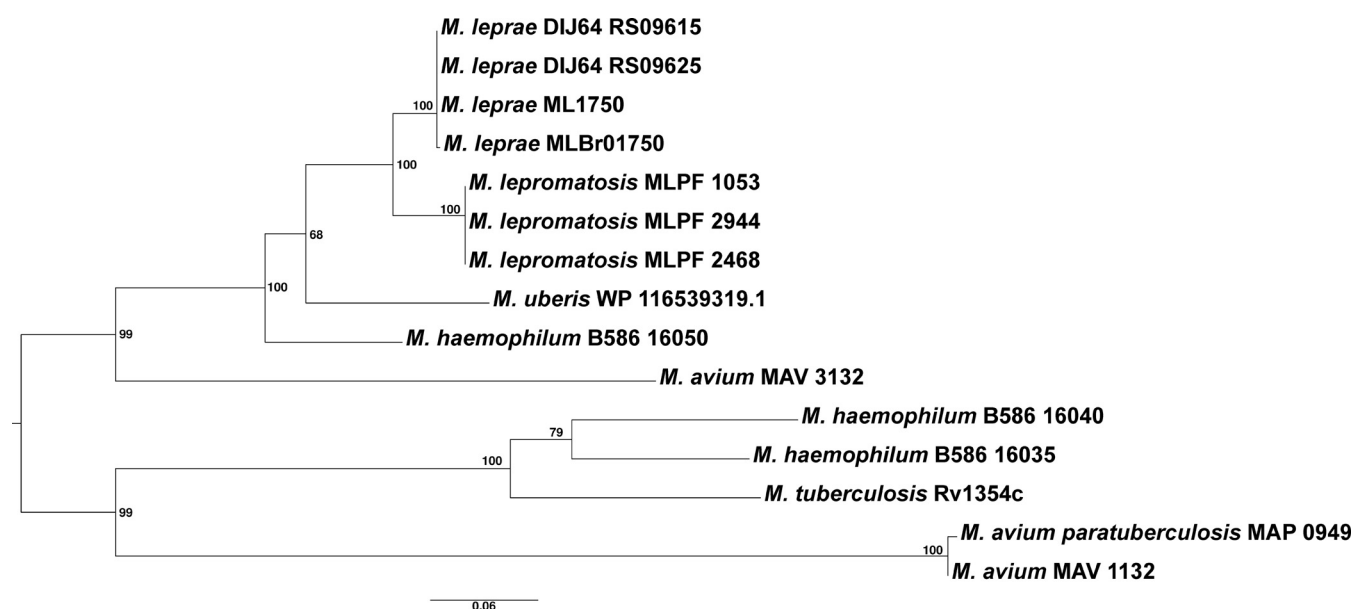
PE/PPE proteins are noted to be antigenic determinants in *M. tuberculosis* to play a role in immuno-pathogenesis (36). To the less virulent leprosy bacilli, these alarming molecules would be detrimental to the survival. Rather, in *M. leprae*, preferential inactivation of PE/PPE genes is noted whereas PGL-I and its genes, beneficial to the bacillus, are retained, which likely aid immune evasion and survival of the bacillus during ~20 million years of reductive evolution (11). In *M. lepromatosis*, these were also true, with statistical significance (Table 3). As a benchmark for gene retention, all 52 genes but one that encode ribosomal proteins, essential for ribosome structure, protein synthesis and bacterial survival, remained (Table 3). The same was true for *M. leprae*. The similar numbers of genes and/or pseudogenes for PE/PPE, PGL-I, and ribosomal proteins suggest their origins in the MRCA and persistence.

Dermal tropism is innate to the clade of *M. haemophilum*, lineages of leprosy bacilli, and *M. ubris*. Despite unknown environmental niches, *M. haemophilum* grows optimally at 30–32°C in culture and requires heme supplement (12, 37). The bacillus causes rare infections, mainly dermal and/or disseminated ones, in immunocompromised patients, such as long-term steroid users. In patients with leprosy, the most common sites of infection involve earlobes, extremities, nasal mucosa, and scrotum where the temperatures tend to be below 37°C to favor bacterial proliferation.

Finally, the large number of shared genes and pseudogenes would require a systematic approach to analyze their functional clusters and/or lack thereof. In view of the cultivation difficulty that has impeded the research for leprosy bacilli, future research endeavors in *M. tuberculosis* and *M. haemophilum* should help in this regard.

**Pathogenetic insights from unique genome features.** The two leprosy bacilli differed in CDS and pseudogenes, including 369 unique CDS in *M. lepromatosis* and 184 unique CDS in *M. leprae* (Table 2). Examination of these unique CDS and pseudogenes should offer insights into some varying pathological features in patients with each infection, such as more skin nodules associated with *M. leprae* but more invasion into subcutis, vessels and internal organs associated with *M. lepromatosis*.

The finding of three copies of the diguanylate cyclase/phosphodiesterase gene in *M. lepromatosis* (MLPF\_1053, 2468 and 2944) contrasted the single copy status in *M. leprae* (MLBr01750), *M. ubris* (WP\_116539319.1) and *M. tuberculosis* (Rv1354c) but resembled those in *M. haemophilum*. A phylogenetic reconstruction of the evolution of these proteins in several mycobacterial species (Fig. 7) revealed that the encoding gene was duplicated in the early ancestor of these species and *Mycobacterium avium* (taxonomy in Fig. 1). After this event, several gene gains and losses took place. *M. lepromatosis*, *M. leprae* and *M. ubris* retained the same copy and lost the other, in opposition to *M. tuberculosis*. Gene MLPF\_1053 was located at a large syntenic block of > 200 kb (block 11, Table S4), suggesting ancestral origin. Gene MLPF\_2468 arose through a 5,020-bp duplication involving MLPF\_1053 and 1052 (a pseudogene derived from the transposase gene of an insertion sequence element) and the last part of pseudogene MLPF\_1051. Gene MLPF\_2944 was a duplicate of a 4,700-bp segment involving MLPF\_2468 and the pseudogenes MLPF\_2467 and 2466. In *M. leprae*, the amplified copies as well as associated pseudogenes were absent (Table S4). Surprisingly, in a recent *M. leprae* genome, MRHRU-235-G from India (CP029543.1, direct submission), a tandem gene duplication was also noted.



**FIG 7** Phylogenetic reconstruction of diguanylate cyclase/phosphodiesterase genes in selected mycobacterial species. Maximum likelihood phylogeny obtained with the amino acid alignment (601 sites), using the evolutionary model JTT+G. Numbers at the nodes indicate bootstrap values obtained with 500 replicates.

The diguanylate cyclase/phosphodiesterase is a well-studied enzyme in diverse bacteria, harboring three domains: GAF as sensor domain, GGDEF as diguanylate cyclase, and EAL as c-di-GMP specific phosphodiesterase (38). The latter are two antagonistic enzymes, hence bifunctional, involved in the synthesis and hydrolysis of the second messenger cyclic di-GMP. The presence of three copies of its gene hints more c-di-GMP metabolism in *M. lepromatosis* for survival and/or pathogenicity in view that deletion of this gene in *M. tuberculosis* affects its dormancy and pathogenicity (39, 40). In *M. leprae*, studies have also suggested a potential role of diguanylate cyclase in the signaling response for intracellular survival (41). In other pathogens, a strategy has been designed to intercept c-di-GMP signaling pathways by directly targeting this second messenger as a way of controlling pathogen survival, growth or biofilm formation (42).

The gene *hemW*, previously annotated as *hemN* (10), was found in *M. lepromatosis* (MLPF\_2234) but not in *M. leprae*. This gene likely encodes a heme chaperone that catalyzes the insertion of heme into hemoproteins for respiration (43, 44). The gene and adjacent MLPF\_2235 (a pseudogene) stood between syntenic blocks 19 and 20 (Table S4); in *M. leprae*, however, the syntenic blocks rearranged afar after divergence, during which *hemW* and adjacent pseudogene (together ~2 kb) were likely lost. In *M. haemophilum* and *M. tuberculosis*, this gene was B586\_09185 and Rv2388c, respectively, and their corresponding amino acids shared 89% and 81% identities with *M. lepromatosis*. This finding warrants further research into the function of *hemW*.

The two leprosy bacilli showed 6 genes in common to encode TetR family transcriptional regulators (*ethR* and others), and there were 5 additional ones in *M. leprae* but only pseudogenes in *M. lepromatosis*. These regulators are involved in diverse functions, such as regulation of c-di-GMP, ethionamide resistance, lipogenesis, etc. (45, 46). The difference in these regulators, along with many other unique genes, suggest considerable variations between the leprosy bacilli in metabolism, stress response, and interaction with the host to influence pathogenesis.

**Conclusions.** The history of the leprosy bacterial lineage started around 30 Mya when a mycobacterial ancestor diverged from the lineage of *M. haemophilum*. At that time, this ancestor would have a similar lifestyle to the present bacterium *M. haemophilum*, perhaps being able to produce opportunistic skin infections (12). Soon after this point, two lineages evolved, one leading to *M. uberis*, a pathogen infecting cows and goats (22) and the other leading to the MRCA of *M. leprae* and *M. lepromatosis*.

Based on the phylogenies and pseudogene comparisons, the lineages of *M. uberis* and the MRCA likely underwent independent processes of reductive genome evolution (22). While the two leprosy lineages diverged around 13.9 Mya (10), our estimation of the ages of pseudogenes suggest that most of them formed shortly before the divergence. Afterwards, additional events of pseudogenization took place independently in both lineages, but at a much lesser scale.

We proposed previously that the reductive evolution of leprosy bacilli began in the MRCA as a parasite within the host, family *Hominidae* (great apes), ~18 to 20 Mya (11). In another word, the specific taming-adapting process has persisted all along until modern humans today. Our current genomic study refines the bacterial reductive evolution to ~14 Mya, which brings the host more specifically within subfamily *Homininae* (humans, gorillas, and chimpanzees). The genome reduction of *M. uberis* that also signifies a host specific process lends further support. Likewise, genome reduction of *Mycobacterium lepraemurium* from ancestral *M. avium* is likely unique to mice in causing murine leprosy (47).

Human *M. leprae* strains have been reported to infect a few animal species, such as nine-banded armadillos in the southern United States (8, 48) and nonhuman primates (48, 49). In the isolated infections of British red squirrels, the *M. leprae* strains were of human origin by hundreds of years while the *M. lepromatosis* strains bore a divergence of ~27,000 years from Mexican strain Mx1-22A (21). In red squirrels from continental northwest Europe, a recent study didn't find such infections (50). Therefore, considering the finding of *M. lepromatosis* in Asian patients with leprosy (17, 18) and the likely introduction of the agent into Mexico 13,000 years ago during the earliest human Asia-Alaska-America migration (14), the most likely source of this bacillus in the British red squirrels could also be migrant humans from Asia ~27,000 years ago. In this regard, future studies on *M. lepromatosis* strains from Asian patients would be confirmative. Together, the new data and literature further corroborate our proposal.

The two leprosy bacilli share 71.4% of the genome but differ in 28.6%. The shared features should account for the dermal and neuronal invasion – the hallmark of leprosy. The preferential retention of PGL-I and its genes, favorable to bacterial survival, but inactivation of the detrimental PE/PPE components, are consistent features. Conversely, the variable genome features, in combination with the host genetic factors in diverse populations, should explain the variations in clinical and pathological manifestations of each infection, such as DLL and vascular occlusion associated with *M. lepromatosis*. Examples include three copies of the bifunctional diguanylate cyclase/phosphodiesterase gene in *M. lepromatosis* but one copy in *M. leprae*, 6 genes encoding the TetR transcriptional regulators in *M. lepromatosis* but 11 such genes in *M. leprae*, the presence of *hemW* in *M. lepromatosis* but absence in *M. leprae*, and so on. Clearly, the large numbers of shared and unique genome features require far more studies to delineate their relations with the infection. To this end, the annotated complete genome of *M. lepromatosis* has laid a foundation. In addition, recent report of passage of *M. lepromatosis* in mouse footpad (16) may provide viable bacteria for more studies as well as possible animal models in the near future.

The present study of the complete genome of *M. lepromatosis* also adds to the general knowledge on reductive genome evolution. This has been well documented in several bacterial endosymbionts and some other pathogens (51–54), involving a shift in lifestyle from relative free-life to host restriction and intracellular living. These genomes undergo chromosomal instability, rearrangements, loss of dispensable genes through either pseudogenization by point mutations or large deletions, and a tendency for genome downsizing. In general, the rates of nucleotide substitution in CDS also increase drastically, except a few examples of low rates in some long-term endosymbionts (55–58). Similar processes of massive pseudogenization have been detected in recent insect-associated bacterial symbionts due to their rich and stable new niches (59–61).

## MATERIALS AND METHODS

**Source of DNA and genome construction.** The *M. lepromatosis* DNA from strain FJ924 was extracted from a smear slide that was prepared from autopsy liver tissue of a patient originally from Mexico (4). The

smear had been dried, acid-fast stained (Kinyoun method) for microscopy, and archived for 6 years (2007–2013). Total genomic DNA was obtained with the QIAamp kit (Qiagen, Valencia, CA) following instructions from manufacturer. The extraction yielded ~3 ng DNA. A whole-genome library was then constructed using the KAPA kit (Kapa Biosystems, Wilmington, MA). The library was enriched by six PCR cycles, quantified, assessed for size distribution, and sequenced on the HiSeq 2000 sequencer with the 75-bp pair end configuration (Illumina, San Diego, CA). A total number of 69 million reads were obtained.

The construction involved several steps. Human DNA contamination was removed first (14 million reads). From the remaining 55 million reads, mycobacterial DNA (11 million reads) were collected through BLAST matches (62) with *M. leprae* Br4923, filtering mostly contaminant DNA (44 million reads) from diverse other bacteria. The specific reads were assembled *de novo* using Velvet v1.2.10 (63), and the assembled contigs were aligned manually as well as using Bowtie v2.1.0 (64) to *M. leprae* template to build the first draft *M. lepromatosis* FJ924 genome. This draft was refined with GapFiller v1-10 (65) for gap closure and extension of contig edges, leading to announcement of a draft genome of 3,215,823 nucleotides with 39 contigs (66).

The published draft genome was used to map reads again, resulting in 12 million reads. These reads were assembled *de novo* with MIRA v2.1 with default parameters (67), and alignment of the assembled contigs led to third draft. The process of mapping, assembly, Gap filling, and draft construction was reiterated 5 times until seal of all blunt-end gaps. Eventually, a high-quality draft genome from 12.3 million reads and 525× genome coverage, with 12 gaps that were all flanked by repetitive contigs, was obtained. Primers were designed at nonrepetitive regions of contig edges and PCRs along with Sanger sequencing of amplicons were used to close these gaps.

**Genome annotation.** The closed *M. lepromatosis* FJ924 genome was annotated independently with Basys (68), RAST web-servers (69), and Prokka v1.12 (70). Annotation results from the three methods were compared and mixed with BEACON v1.1 (71). Initial pseudogene prediction was based on Prokka annotation of adjacent CDS presenting the same annotation and a custom python script. In brief, the script uses LAST (72) alignments to detect CDS which are 75% shorter than a reference set of clustered proteomes at 95% amino acid identity (Table S8). Then, it mixes adjacent fragmented genes when they overlap with the same reference hit.

Pseudogenes annotations were revised exhaustively by using several types of BLAST analyses. The main one was based on BLASTX using the complete *M. lepromatosis* FJ924 genome sequence as a query against a protein database of *M. haemophilum* (E value 0.01). This revision produced length changes of several pseudogenes, fusion of others, and conversion of some initially annotated CDS to pseudogenes.

**Comparative analysis of *M. lepromatosis* strains FJ924 and Mx1-22A.** To examine the differences between the draft genome Mx1-22A and the complete genome of strain FJ924, the 126 contigs of Mx1-22A were queried to FJ924 in BLASTN searches (E value E-180). The results produced a complete alignment of 134 shared segments. The noncovered segments of FJ924 were identified and their annotation features were extracted (for those higher than 300 bp) (Table S1). SNPs between the two strains were identified with MAUVE (23). The initial SNP number of 314 was reduced to 160 after removing ambiguous nucleotides in the contigs of Mx1-22A. In the second round, all inaccuracies and SNPs within the end 500 nucleotides of the contigs were removed, leaving behind a set of 17 SNPs. These SNPs were checked, and six of them were removed due to their location in a short segment with many ambiguous nucleotides. A final set of 11 SNPs were studied.

**Phylogenetics.** Several available mycobacterial proteomes were selected for phylogenetic reconstruction. OrthoFinder v2.3.3 (27) was used to detect, align, and concatenate 1,220 single core orthologous clusters of proteins shared between *M. lepromatosis* FJ924, *M. leprae* Br4923, *M. haemophilum* DSM44634, *M. tuberculosis* H37Rv and variant AF2122/97, *M. ubris* Jura, *M. marinum* M, *M. ulcerans* SGL03, *M. avium* 104 and subsp. *paratuberculosis* K-10 (mafft aligner and IQ-TREE options) (27, 73, 74). Obtained concatenated alignment was pruned with Gblocks v0.91b (75) with the option *no gaps allowed*. A total of 316,716 amino acid positions from 418,308 were maintained and used for phylogenetic reconstruction. The species tree was inferred by the Maximum Likelihood method (JTT+I+F+R3 model) using IQ-TREE v1.6.12 with 1000 ultrafast bootstrap and SH-aLRT as node support (74). An additional phylogenetic tree was inferred by the Maximum Likelihood method (JTT+G+I+F model and 500 bootstrap replicates) for 15 diguanylate cyclase/phosphodiesterase proteins of several mycobacterial species (601 sites) in MEGA7 (76).

A time tree was inferred by applying the RelTime method (77) using the previous concatenated alignment and phylogenetic tree (316,716 sites and 10 sequences), a model JTT+G+I+F and branch lengths estimation by the ordinary least-squares approach in MEGA11 (78). The calibration constraint was the minimum and maximum time of the node of divergence between the two leprosy bacilli (8.2–21.4 Mya) as previously reported (10).

**Syntenic and chromosomal rearrangement analysis.** Alignment of several mycobacterial genomes was performed with MAUVE (23) using default parameters. GenoPlotR was used to plot Mauve results (79). The permutation file was converted with the program GRIMM in a distant matrix, which contained the minimal number of inversions required between a pair of genomes to explain differences in gene order (26). Distance matrix was charged in MEGA7 (76) and a neighbor joining algorithm was used to infer the rearrangement phylogeny (80). Orthology between CDS and pseudogenes from *M. leprae* Br4923 and *M. lepromatosis* FJ924 was obtained with a reciprocal BLASTN best hit strategy (E value = 1.0E-05). Nonreciprocal hits and those hits in disagreement with gene order were revised upon visual inspection. Synteny was the first criterion to assign orthology when more than one similar hit was detected allowing up to 5 hits in new BLASTN searches.

**Identification and analysis of repetitive sequences.** A BLASTN search using the genome of *M. lepromatosis* as both query and subject with an E value of  $10^{-20}$  was performed to search for repetitive



sequences. Segments with more than one overlapping hit were revised to produce the repeat segment with the largest length. Repetitive sequences of the same family were aligned in Ugene v33.0 (81). To determine the effect of repeat segments over chromosomal rearrangements, the 24 synteny breaks were delimited by the last CDS/pseudogene in a block and the first CDS/pseudogene in the next block with an orthologous in the genome of *M. leprae*. A plot comparing repetitive sequences and synteny breaks was made with Circos (82).

**Identification of orthologous clusters.** Orthologous clusters of proteins (genes) in the forms of core, pan, pairwise shared, and strain specific clusters were obtained with OrthoFinder v2.3.3 using the *msa* (mafft aligner) and IQ-TREE options (27, 73, 74). Data on the clusters were used for several analyses.

**Estimation of the ages of pseudogenes.** The relative ages of *M. leprae* and *M. lepromatosis* pseudogenes were estimated according to the method of Gomez-Valero et al. (7) with inclusion of the recent *M. haemophilum* genome data (12) for closer phylogeny. Briefly, this method used *M. tuberculosis* as out-group to root the tree and compared *M. haemophilum* with either *M. leprae* or *M. lepromatosis*. The divergence between the two species determines the relative age of pseudogenization, ranging from 1, when it took place at the start of the divergence, to 0 at this time. The method is based on the estimation of the dN and dS (the number of nonsynonymous and synonymous substitutions per site, respectively) for each cluster of orthologous gene-pseudogene in the branches leading to *M. haemophilum* and a leprosy bacillus.

Estimation of the dN and dS values required use of clusters of orthologous genes. To produce clusters of orthology that incorporated *M. leprae* and/or *M. lepromatosis* pseudogenes and corresponding genes from both *M. haemophilum* and *M. tuberculosis*, a reciprocal best hit strategy was used with E value 1.0E-05. The set of pseudogenes/genes of *M. lepromatosis* and *M. leprae* were then compared, and they were further aligned to the sets of *M. haemophilum* and *M. tuberculosis*.

Alignments of genes and pseudogenes to maintain the codon structure, despite the indel mutations of the pseudogenes, were performed with MACSE (83). This program was also used to align clusters of functional genes in the four species. The program codeML (84) was used to obtain the dN and dS values as previously described (58). A custom python script was used to run those analyses.

To estimate relative pseudogene ages (*Rpa*), the following formulas were used, for *M. lepromatosis* (up) and for *M. leprae* (down):

$$Rpa = (dN_{ilp_{ps}} - f(dN_{ih})) / (\overline{dS}_{ilp_{ps}} - f(dN_{ih}))$$

$$Rpa = (dN_{ile_{ps}} - f'(dN_{ih})) / (\overline{dS}_{ile_{ps}} - f'(dN_{ih}))$$

The *i* (intercept) corresponds to the node of divergence between the leprosy bacilli and *M. haemophilum* lineages. The values  $dN_{ilp_{ps}}$  and  $dN_{ile_{ps}}$  are the numbers of nonsynonymous substitutions in the branches of a pseudogene from *M. lepromatosis* and *M. leprae*, respectively. To be pointed out, the term dN only shows a strict sense for the period of evolution as a gene.

The parameters  $f(dN_{ih})$  and  $f'(dN_{ih})$  are estimates, for orthologous genes, of the expected dN value for the branch of a leprosy bacillus as a function of the dN value for the branch of *M. haemophilum* considering that this gene had been evolving as a functional gene for the complete period of time. Orthologous genes evolve faster in the leprosy bacilli than in *M. haemophilum*, especially for dN (four times faster on average). By using dN values of 1,213 clusters of orthologous genes, the relationship between outcome and predictor, both log-transformed, was estimated by segmented linear regression. A two-piece linear relationship was assumed, namely, represented by two straight lines connected at an unknown breakpoint. Estimation was performed by means of Muggeo's algorithm (85) as implemented in the library *segmented* of the R software (version 4.0.4) (86). The last term of the formula, the average dS in the branch leading to *M. lepromatosis* and *M. leprae* pseudogenes, may be considered a value close to the number of substitutions per site in a nonfunctional DNA sequence (0.2851 and 0.2813 for *M. lepromatosis* and *M. leprae*, respectively). These are the expected dN values for pseudogenes with *Rpa* values of 1.

The time of divergence in Mya between the lineage of leprosy bacilli and the *M. haemophilum* lineage (*Rpa* = 1) was estimated with the RelTime method (28). All mentioned scripts are available at <https://github.com/DiegoSantos-Garcia/Scripts>.

**Data availability.** This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number CP083405 and BioProject PRJNA281005.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**SUPPLEMENTAL FILE 1**, XLSX file, 0.02 MB.

**SUPPLEMENTAL FILE 2**, XLSX file, 0.01 MB.

**SUPPLEMENTAL FILE 3**, XLSX file, 0.02 MB.

**SUPPLEMENTAL FILE 4**, XLSX file, 0.4 MB.

**SUPPLEMENTAL FILE 5**, XLSX file, 0.01 MB.

**SUPPLEMENTAL FILE 6**, XLSX file, 0.01 MB.

**SUPPLEMENTAL FILE 7**, XLSX file, 0.02 MB.

**SUPPLEMENTAL FILE 8**, XLSX file, 0.01 MB.

## ACKNOWLEDGMENTS

This work was supported in part by a University Cancer Foundation grant from the University of Texas MD Anderson Cancer Center (to X.Y.H.), by the National Institutes of Health (NIH) grant P30 CA016672 for MD Anderson's core Sequencing and Microarray Facility and core Bioinformatics Shared Resource, by the NIH grant CA143883 for MD Anderson's TCGA Genome Data Analysis Center, and by the Mary K. Chapman Foundation. D.S.G. was recipient of a contract from the project ANR Hmicmac 16-CE02-0014. F.J.S. was funded by Ministerio de Ciencia, Innovación y Universidades (Spain) (PGC2018-099344-B-I00) and Generalitat Valenciana (Prometeo/2018/A/133) and cofinanced by the European Regional Development Fund (ERDF). We thank Nipun Mistry for participation in initial work. We also thank Carmen Íñiguez for her support in segmented linear regression estimations.

X.Y.H., X.Z., and L.Z. performed genome assembly. F.J.S. and D.S.G. performed annotation and data analyses. F.J.S. and X.Y.H. drafted the manuscript, and all authors edited and approved of the manuscript.

## REFERENCES

1. Gelber RH. 2005. Leprosy (Hansen's Disease). P 966–972. In Kasper DL, Braunwald E, Fauci AS, Hauser SL, Longo DL, Jameson JL (ed), Harrison's principles of internal medicine, 16th ed McGraw-Hill, New York.
2. Latapi F, Chevez Zamora A. 1948. The "spotted" leprosy of Lucio: an introduction to its clinical and histological study. *Int J Lepr* 16:421–430.
3. Hansen GHA. 1874. Undersøgelser angående spedalskhedens årsager. *Nor Mag Laegevidenskaben* 9:1–88.
4. Han XY, Seo Y-H, Sizer KC, Schoberle T, May GS, Spencer JS, Li W, Nair RG. 2008. A new *Mycobacterium* species causing diffuse lepromatous leprosy. *Am J Clin Pathol* 130:856–864. <https://doi.org/10.1309/AJCPP72FJZZRRVMM>.
5. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honoré N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG. 2001. Massive gene decay in the leprosy bacillus. *Nature* 409:1007–1011. <https://doi.org/10.1038/35059006>.
6. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekai F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393:537–544. <https://doi.org/10.1038/31159>.
7. Gomez-Valero L, Rocha EPC, Latorre A, Silva FJ. 2007. Reconstructing the ancestor of *Mycobacterium leprae*: the dynamics of gene loss and genome reduction. *Genome Res* 17:1178–1185. <https://doi.org/10.1101/gr.6360207>.
8. Monot M, Honoré N, Garnier T, Araoz R, Coppée JY, Lacroix C, Sow S, Spencer JS, Truman RW, Williams DL, Gelber R, Virmond M, Flageul B, Cho SN, Ji B, Paniz-Mondolfi A, Convit J, Young S, Fine PE, Rasolofso V, Brennan PJ, Cole ST. 2005. On the origin of leprosy. *Science* 308:1040–1042. <https://doi.org/10.1126/science.1109759>.
9. Han XY, Sizer KC, Thompson EJ, Kabanja J, Li J, Hu P, Gómez-Valero L, Silva FJ. 2009. Comparative sequence analysis of *Mycobacterium leprae* and the new leprosy-causing *Mycobacterium lepromatosis*. *J Bacteriol* 191:6067–6074. <https://doi.org/10.1128/JB.00762-09>.
10. Singh P, Benjak A, Schuenemann VJ, Herbig A, Avanzi C, Busso P, Nieselt K, Krause J, Vera-Cabrera L, Cole ST. 2015. Insight into the evolution and origin of leprosy bacilli from the genome sequence of *Mycobacterium lepromatosis*. *Proc Natl Acad Sci U S A* 112:4459–4464. <https://doi.org/10.1073/pnas.1421504112>.
11. Han XY, Silva FJ. 2014. On the age of leprosy. *PLoS Negl Trop Dis* 8:e2544. <https://doi.org/10.1371/journal.pntd.0002544>.
12. Tufariello JM, Kerantzas CA, Vilchère C, Calder RB, Nordberg EK, Fischer JA, Hartman TE, Yang E, Driscoll T, Cole LE, Sebra R, Maqbool SB, Wattam AR, Jacobs WR. 2015. The complete genome sequence of the emerging pathogen *Mycobacterium haemophilum* explains its unique culture requirements. *mBio* 6:e01313-13115–e01315. <https://doi.org/10.1128/mBio.01313-15>.
13. Vera-Cabrera L, Escalante-Fuentes WG, Gomez-Flores M, Ocampo-Candiani J, Busso P, Singh P, Cole ST. 2011. Case of diffuse lepromatous leprosy associated with "*Mycobacterium lepromatosis*". *J Clin Microbiol* 49:4366–4368. <https://doi.org/10.1128/JCM.05634-11>.
14. Han XY, Sizer KC, Velarde-Félix JS, Frias-Castro LO, Vargas-Ocampo F. 2012. The leprosy agents *Mycobacterium lepromatosis* and *Mycobacterium leprae* in Mexico. *Int J Dermatol* 51:952–959. <https://doi.org/10.1111/j.1365-4632.2011.05414.x>.
15. Jessamine PG, Desjardins M, Gillis T, Scollard D, Jamieson F, Broukhanski G, Chedore P, McCarthy A. 2012. Leprosy-like illness in a patient with *Mycobacterium lepromatosis* from Ontario. *Canada J Drugs Dermatol* 11:229–233.
16. Sharma R, Singh P, McCoy RC, Lenz SM, Donovan K, Ochoa MT, Estrada-Garcia I, Silva-Miranda M, Jurado-Santa Cruz F, Balagon MF, Stryjewska B, Scollard DM, Pena MT, Lahiri R, Williams DL, Truman RW, Adams LB. 2020. Isolation of *Mycobacterium lepromatosis* and development of molecular diagnostic assays to distinguish *Mycobacterium leprae* and *M. lepromatosis*. *Clin Infect Dis* 71:e262–e269. <https://doi.org/10.1093/cid/ciz1121>.
17. Han XY, Aug FM, Choon SE, Werner B. 2014. Analysis of the leprosy agents *Mycobacterium leprae* and *Mycobacterium lepromatosis* in four countries. *Am J Clin Pathol* 142:524–532. <https://doi.org/10.1309/AJCP1GLCBE5CDZRM>.
18. Han XY, Sizer KC, Tan H-H. 2012. Identification of the leprosy agent *Mycobacterium lepromatosis* in Singapore. *J Drugs Dermatol* 11:168–172.
19. Rea TH, Jersey RS. 2005. Clinical and histologic variations among thirty patients with Lucio's phenomenon and pure and primitive diffuse lepromatosis (Latapi's lepromatosis). *Int J Lepr Other Mycobact Dis* 73:169–188.
20. Vargas-Ocampo F. 2007. Diffuse leprosy of Lucio and Latapi: a histologic study. *Lepr Rev* 78:248–260. <https://doi.org/10.47276/lr.78.3.248>.
21. Avanzi C, Del-Pozo J, Benjak A, Stevenson K, Simpson VR, Busso P, McLuckie J, Loiseau C, Lawton C, Schoening J, Shaw DJ, Piton J, Vera-Cabrera L, Velarde-Felix JS, McDermott F, Gordon SV, Cole ST, Meredith AL. 2016. Red squirrels in the British Isles are infected with leprosy bacilli. *Science* 354:744–747. <https://doi.org/10.1126/science.aah3783>.
22. Benjak A, Avanzi C, Benito Y, Breyse F, Chartier C, Boschirollo M-L, Fourichon C, Michelet L, Pin D, Flandrois J-P, Bruyere P, Dumitrescu O, Cole ST, Lina G. 2018. Highly reduced genome of the new species *Mycobacterium uberis*, the causative agent of nodular thelitis and tuberculoid scrotoitis in livestock and a close relative of the leprosy bacilli. *mSphere* 3:e00405-18. <https://doi.org/10.1128/mSphere.00405-18>.
23. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14:1394–1403. <https://doi.org/10.1101/gr.2289704>.
24. Darling AE, Tritt A, Eisen JA, Facciotti MT. 2011. Mauve assembly metrics. *Bioinformatics* 27:2756–2757. <https://doi.org/10.1093/bioinformatics/btr451>.
25. Cole ST, Sizer KC, Honoré N. 2001. Repetitive sequences in *Mycobacterium leprae* and their impact on genome plasticity. *Lepr Rev* 72:449–461.
26. Tesler G. 2002. GRIMM: genome rearrangements web server. *Bioinformatics* 18:492–493. <https://doi.org/10.1093/bioinformatics/18.3.492>.

27. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20:238. <https://doi.org/10.1186/s13059-019-1832-y>.
28. Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. *Proc Natl Acad Sci U S A* 109:19333–19338. <https://doi.org/10.1073/pnas.1213199109>.
29. Rambukkana A. 2001. Molecular basis for the peripheral nerve predilection of *Mycobacterium leprae*. *Curr Opin Microbiol* 4:21–27. [https://doi.org/10.1016/s1369-5274\(00\)00159-4](https://doi.org/10.1016/s1369-5274(00)00159-4).
30. Mehra V, Brennan PJ, Rada E, Convit J, Bloom BR. 1984. Lymphocyte suppression in leprosy induced by unique *M. leprae* glycolipid. *Nature* 308:194–196. <https://doi.org/10.1038/308194a0>.
31. Tabouret G, Astarie-Dequeker C, Demangel C, Malaga W, Constant P, Ray A, Honoré N, Bello NF, Perez E, Daffé M, Guilhot C. 2010. *Mycobacterium leprae* phenolglycolipid-1 expressed by engineered *M. bovis* BCG modulates early interaction with human phagocytes. *PLoS Pathog* 6:e1001159. <https://doi.org/10.1371/journal.ppat.1001159>.
32. Enany S, Yoshida Y, Tateishi Y, Ozeki Y, Nishiyama A, Savitskaya A, Yamaguchi T, Ohara Y, Yamamoto T, Ato M, Matsumoto S. 2017. Mycobacterial DNA-binding protein 1 is critical for long term survival of *Mycobacterium smegmatis* and simultaneously coordinates cellular functions. *Sci Rep* 7:6810. <https://doi.org/10.1038/s41598-017-06480-w>.
33. Savitskaya A, Nishiyama A, Yamaguchi T, Tateishi Y, Ozeki Y, Nameta M, Kon T, Kaboso SA, Ohara N, Peryanova OV, Matsumoto S. 2018. C-terminal intrinsically disordered region-dependent organization of the mycobacterial genome by a histone-like protein. *Sci Rep* 8:8197. <https://doi.org/10.1038/s41598-018-26463-9>.
34. Gröschel MI, Sayes F, Simeone R, Majlessi L, Brosch R. 2016. ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat Rev Microbiol* 14:677–691. <https://doi.org/10.1038/nrmicro.2016.131>.
35. Roy S, Ghatak D, Das P, BoseDasgupta S. 2020. ESX secretion system: the gatekeepers of mycobacterial survivability and pathogenesis. *Eur J Microbiol Immunol (Bp)* 10:202–209. <https://doi.org/10.1556/1886.2020.00028>.
36. Brennan MJ. 2017. The enigmatic PE/PPE multigene family of mycobacteria and tuberculosis vaccination. *Infect Immun* 85:e00969-16. <https://doi.org/10.1128/IAI.00969-16>.
37. Saubolle MA, Kiehn TE, White MH, Rudinsky MF, Armstrong D. 1996. *Mycobacterium haemophilum*: microbiology and expanding clinical and geographic spectra of disease in humans. *Clin Microbiol Rev* 9:435–447. <https://doi.org/10.1128/CMR.9.4.435>.
38. Galperin MY. 2018. What bacteria want. *Environ Microbiol* 20:4221–4229. <https://doi.org/10.1111/1462-2920.14398>.
39. Bharati BK, Sharma IM, Kasetty S, Kumar M, Mukherjee R, Chatterji D. 2012. A full-length bifunctional protein involved in c-di-GMP turnover is required for long-term survival under nutrient starvation in *Mycobacterium smegmatis*. *Microbiology (Reading)* 158:1415–1427. <https://doi.org/10.1099/mic.0.053892-0>.
40. Hong Y, Zhou X, Fang H, Yu D, Li C, Sun B. 2013. Cyclic di-GMP mediates *Mycobacterium tuberculosis* dormancy and pathogenicity. *Tuberculosis (Edinb)* 93:625–634. <https://doi.org/10.1016/j.tube.2013.09.002>.
41. Rotcheewaphan S, Belisle JT, Webb KJ, Kim H-J, Spencer JS, Borlee BR. 2016. Diguanylate cyclase activity of the *Mycobacterium leprae* T cell antigen ML1419C. *Microbiology (Reading)* 162:1651–1661. <https://doi.org/10.1099/mic.0.000339>.
42. Hee C-S, Habazettl J, Schmutz C, Schirmer T, Jenal U, Grzesiek S. 2020. Intercepting second-messenger signaling by rationally designed peptides sequestering c-di-GMP. *Proc Natl Acad Sci U S A* 117:17211–17220. <https://doi.org/10.1073/pnas.2001232117>.
43. Sharma M, Gupta Y, Dwivedi P, Kempaiah P, Singh P. 2021. *Mycobacterium lepromatosis* MLPM\_5000 is a potential heme chaperone protein HemW and mis-annotation of its orthologues in mycobacteria. *Infect Genet Evol* 94:105015. <https://doi.org/10.1016/j.meegid.2021.105015>.
44. Haskamp V, Karrie S, Mingers T, Barthels S, Alberge F, Magalon A, Müller K, Bill E, Lubitz W, Kleeberg K, Schweyen P, Bröring M, Jahn M, Jahn D. 2018. The radical SAM protein HemW is a heme chaperone. *J Biol Chem* 293:2558–2572. <https://doi.org/10.1074/jbc.RA117.000229>.
45. Deng W, Li C, Xie J. 2013. The underlying mechanism of bacterial TetR/AcrR family transcriptional repressors. *Cell Signal* 25:1608–1613. <https://doi.org/10.1016/j.cellsig.2013.04.003>.
46. Ramos JL, Martínez-Bueno M, Molina-Henares AJ, Terán W, Watanabe K, Zhang X, Gallegos MT, Brennan R, Tobes R. 2005. The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev* 69:326–356. <https://doi.org/10.1128/MMBR.69.2.326-356.2005>.
47. Benjak A, Honap TP, Avanzi C, Becerril-Villanueva E, Estrada-García I, Rojas-Espinoza O, Stone AC, Cole ST. 2017. Insights from the genome sequence of *Mycobacterium lepraemurium*: massive gene decay and reductive evolution. *mBio* 8:e01283-17. <https://doi.org/10.1128/mBio.01283-17>.
48. Ploemacher T, Faber WR, Menke H, Rutten V, Pieters T. 2020. Reservoirs and transmission routes of leprosy; A systematic review. *PLoS Negl Trop Dis* 14:e0008276. <https://doi.org/10.1371/journal.pntd.0008276>.
49. Honap TP, Pfister LA, Housman G, Mills S, Tarara RP, Suzuki K, Cuzzo FP, Sauter ML, Rosenberg MS, Stone AC. 2018. *Mycobacterium leprae* genomes from naturally infected nonhuman primates. *PLoS Negl Trop Dis* 12:e0006190. <https://doi.org/10.1371/journal.pntd.0006190>.
50. Tió-Coma M, Sprong H, Kik M, Dissel JT, Han XY, Pieters T, Geluk A. 2020. Lack of evidence for the presence of leprosy bacilli in red squirrels from North-West Europe. *Transbound Emerg Dis* 67:1032–1034. <https://doi.org/10.1111/tbed.13423>.
51. Murray GGR, Charlesworth J, Miller EL, Casey MJ, Lloyd CT, Gottschalk M, Tucker AW, Welch JJ, Weinert LA. 2021. Genome reduction is associated with bacterial pathogenicity across different scales of temporal and ecological divergence. *Mol Biol Evol* 38:1570–1579. <https://doi.org/10.1093/molbev/msaa323>.
52. Silva FJ, Latorre A, Moya A. 2001. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. *Trends Genet* 17:615–618. [https://doi.org/10.1016/s0168-9525\(01\)02483-0](https://doi.org/10.1016/s0168-9525(01)02483-0).
53. van Ham RCHJ, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M, Silva FJ, Tamames J, Viguera E, Latorre A, Valencia A, Morán F, Moya A. 2003. Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci U S A* 100:581–586. <https://doi.org/10.1073/pnas.0235981100>.
54. Wolf YI, Koonin EV. 2013. Genome reduction as the dominant mode of evolution. *Bioessays* 35:829–837. <https://doi.org/10.1002/bies.201300037>.
55. Bennett GM, Abbà S, Kube M, Marzachi C. 2016. Complete genome sequences of the obligate symbionts “*Candidatus* Sulcia muelleri” and “*Ca.* Nasuia deltocephalinicola” from the pestiferous leafhopper *Macrostelus quadripunctatus* (Hemiptera: Cicadellidae). *Genome Announc* 4:4–5. <https://doi.org/10.1128/genomeA.01604-15>.
56. Silva FJ, Santos-García D. 2015. Slow and fast evolving endosymbiont lineages: positive correlation between the rates of synonymous and non-synonymous substitution. *Front Microbiol* 6:1279. <https://doi.org/10.3389/fmicb.2015.01279>.
57. Santos-García D, Vargas-Chavez C, Moya A, Latorre A, Silva FJ. 2015. Genome evolution in the primary endosymbiont of whiteflies sheds light on their divergence. *Genome Biol Evol* 7:873–888. <https://doi.org/10.1093/gbe/evv038>.
58. Santos-García D, Silva FJ, Morin S, Dettner K, Kuechler SM. 2017. The all-rounder *Sodalis*: a new bacteriome-associated endosymbiont of the lygaeoid bug *Henestaris halophilus* (Heteroptera: Henestarinae) and a critical examination of its evolution. *Genome Biol Evol* 9:2893–2910. <https://doi.org/10.1093/gbe/evx202>.
59. Oakeson KF, Gil R, Clayton AL, Dunn DM, von Niederhausern AC, Hamil C, Aoyagi A, Duval B, Baca A, Silva FJ, Vallier A, Jackson DG, Latorre A, Weiss RB, Heddi A, Moya A, Dale C. 2014. Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biol Evol* 6:76–93. <https://doi.org/10.1093/gbe/evt210>.
60. Toh H, Weiss BL, Perkin SAH, Yamashita A, Oshima K, Hattori M, Aksoy S. 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res* 16:149–156. <https://doi.org/10.1101/gr.4106106>.
61. Santos-García D, Juravel K, Freilich S, Zchori-Fein E, Latorre A, Moya A, Morin S, Silva FJ. 2018. To B or not to B: comparative genomics suggests *Arsenophonus* as a source of B vitamins in whiteflies. *Front Microbiol* 9:2254. <https://doi.org/10.3389/fmicb.2018.02254>.
62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
63. Zerbino DR. 2010. Using the velvet de novo assembler for short-read sequencing technologies. Current protocols in bioinformatics/editorial board, Andreas D. Baxevanis. [et al].
64. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>.
65. Boetzer M, Pirovano W. 2012. Toward almost closed genomes with Gap-Filler. *Genome Biol* 13:R56. <https://doi.org/10.1186/gb-2012-13-6-r56>.
66. Han XY, Mistry NA, Thompson EJ, Tang H-L, Khanna K, Zhang L. 2015. Draft genome sequence of new leprosy agent *Mycobacterium lepromatosis*. *Genome Announc* 3:e00513–515. <https://doi.org/10.1128/genomeA.00513-15>.

67. Chevreaux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. *Comput Sci Biol Proc Ger Conf Bioinforma* 99:45–56.
68. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS. 2005. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res* 33:W455–459. <https://doi.org/10.1093/nar/gki593>.
69. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. <https://doi.org/10.1186/1471-2164-9-75>.
70. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
71. Kalkatawi M, Alam I, Bajic VB. 2015. BEACON: automated tool for Bacterial GEnome Annotation Comparison. *BMC Genomics* 16:616. <https://doi.org/10.1186/s12864-015-1826-4>.
72. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* 21:487–493. <https://doi.org/10.1101/gr.113985.110>.
73. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059–3066. <https://doi.org/10.1093/nar/gkf436>.
74. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–274. <https://doi.org/10.1093/molbev/msu300>.
75. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552. <https://doi.org/10.1093/oxfordjournals.molbev.a026334>.
76. Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874. <https://doi.org/10.1093/molbev/msw054>.
77. Tamura K, Tao Q, Kumar S. 2018. Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates. *Mol Biol Evol* 35:1770–1782. <https://doi.org/10.1093/molbev/msy044>.
78. Tamura K, Stecher G, Kumar S. 2021. MEGA11: molecular evolutionary genetics analysis version 11. *Mol Biol Evol* 38:3022–3027. <https://doi.org/10.1093/molbev/msab120>.
79. Guy L, Kultima JR, Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 26:2334–2335. <https://doi.org/10.1093/bioinformatics/btq413>.
80. Belda E, Moya A, Silva FJ. 2005. Genome rearrangement distances and gene order phylogeny in  $\gamma$ -proteobacteria. *Mol Biol Evol* 22:1456–1467. <https://doi.org/10.1093/molbev/msi134>.
81. Okonechnikov K, Golosova O, Fursov M, UGENE team. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28:1166–1167. <https://doi.org/10.1093/bioinformatics/bts091>.
82. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* 19:1639–1645. <https://doi.org/10.1101/gr.092759.109>.
83. Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol Biol Evol* 35:2582–2584. <https://doi.org/10.1093/molbev/msy159>.
84. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591. <https://doi.org/10.1093/molbev/msm088>.
85. Muggeo VMR. 2003. Estimating regression models with unknown break-points. *Stat Med* 22:3055–3071. <https://doi.org/10.1002/sim.1545>.
86. R Development Core Team, R Core Team 2021, R Core Team. 2021. R: a Language and Environment for Statistical Computing. Vienna, Austria.