A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region

Authors:

Jean Claude Semuto Ngabonziza[1,2,3#], Chloé Loiseau[4,5#], Michael Marceau[6#], Agathe Jouet[7], Fabrizio Menardo[4,5], Oren Tzfadia[2], Esdras Belamo Niyigena[1], Wim Mulders[2], Kristina Fissette[2], Maren Diels[8], Cyril Gaudin[7], Stéphanie Duthoy[7], Willy Ssengooba[9], Emmanuel André[10], Michel K Kaswa[11], Yves Mucyo Habimana[12], Daniela Brites[4,5], Dissou Affolabi[13], Jean Baptiste Mazarati[14], Bouke Catherine de Jong[2], Leen Rigouts[2,3], Sebastien Gagneux[4,5*], Conor Joseph Meehan[2,15*], Philip Supply[6*]

Affiliations:

[1]National Reference Laboratory Division, Department of Biomedical Services, Rwanda Biomedical Center, Kigali, Rwanda. [2]Mycobacteriology Unit, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium. [3]Department of Biomedical Sciences, Antwerp University, Antwerp, Belgium. [4]Swiss Tropical and Public Health Institute, Basel, Switzerland. [5]University of Basel, Basel, Switzerland. [6]Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000 Lille, France. [7]Genoscreen, Lille, France. [8]BCCM/ITM, Mycobacterial culture collection, Institute of Tropical Medicine, Antwerp, Belgium. [9]Department of Medical Microbiology, College of Health Sciences, Makerere University, Kampala, Uganda. [10]Laboratory of Clinical Bacteriology and Mycology, Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium. [11]National Tuberculosis Program, Kinshasa, DR Congo. [12]Tuberculosis and Other Respiratory Diseases Division, Institute of HIV/AIDS Disease Prevention and Control, Rwanda Biomedical Center, Kigali, Rwanda. [13]Laboratoire de Référence des Mycobactéries, Cotonou, Benin. [14]Department of Biomedical Services, Rwanda Biomedical Center. [15]School of Chemistry and Biosciences, University of Bradford, Bradford, UK.

[#]contributed equally

[*]Joint senior authors

Materials and correspondence: Conor Meehan: c.meehan2@bradford.ac.uk, Philip Supply: Philip.supply@ibl.cnrs.fr, Sebastien Gagneux: sebastien.gagneux@swisstph.ch

Abstract

The human- and animal-adapted lineages of the *Mycobacterium tuberculosis* complex (MTBC) are thought to have clonally expanded from a common progenitor in Africa. However, the molecular events that accompanied this emergence remain largely unknown. Here, we describe two MTBC strains isolated from patients with multidrug-resistant tuberculosis, representing an as-yet-unknown lineage, named Lineage 8 (L8), restricted to the African Great Lakes region. Using genome-based phylogenetic reconstruction, we show that L8 is a sister clade to the known MTBC lineages. Comparison with other complete mycobacterial genomes indicate that the divergence of L8 preceded the loss of the *cobF* genome region - involved in the cobalamin/vitamin B12 synthesis - and gene interruptions in a subsequent common ancestor shared by all other known MTBC lineages. This discovery further supports an East African origin for the MTBC and provides additional molecular clues on the ancestral genome reduction associated with adaptation to a pathogenic lifestyle.

Introduction

Tuberculosis (TB), caused by members of the *Mycobacterium tuberculosis* complex (MTBC), is among the most ancient scourges of humankind[1], and remains the leading cause of mortality globally due to an infectious disease[2]. Intense research has been dedicated to decipher the evolutionary history of the MTBC and to understand the causes underlying the worldwide spread of TB[3–5]. Current genome data show that the MTBC is comprised of two main branches, one including the five human-adapted lineages representing *M. tuberculosis sensu stricto* (L1–4, and L7), and the second branch comprising two other human-adapted lineages traditionally referred to as *M. africanum* (L5-6) and at least nine animal-adapted lineages[6]. Africa is the only continent where all MTBC lineages are present, suggesting that the MTBC emerged from a common ancestor therein and then clonally expanded to the rest of the world following human migrations[3,7,8]. However, the genomic traits of this common ancestor and the region from which this expansion took place in Africa remain unknown. Whole genome sequencing (WGS) analyses showed that rare human TB bacilli with a smooth colony morphotype, highly restricted to the Horn of Africa and named *Mycobacterium canettii* (alias smooth tubercle bacilli or STB) represent early evolutionary branching lineages that predate the emergence of the most recent common ancestor (MRCA) of the MTBC (or of the rest of the MTBC, if *M. canettii* is considered to be part of the complex)[4,9,10] Indeed, whereas known MTBC strains differ by no more than ~2,000 Single Nucleotide Polymorphisms (SNPs)[11], *M. canettii* strains are 10 to 25-fold more genetically diverse and separated by at least 14,000 SNPs from the hitherto known MTBC MRCA[4,10]. Moreover, *M. canettii* strains are less virulent and possess highly mosaic genomes, possibly reflecting primal adaptation to an environmental reservoir favouring active lateral gene flow[4,12,13]. These biological differences support the existence of intermediate lineages in the evolution from *M. canettii* towards the obligate MTBC pathogens.

Here, we describe two exceptional strains representing a new, third main branch, diverging before the MRCA of the other MTBC lineages. These two strains, isolated in Rwanda and Uganda, respectively, were discovered in two independent analyses, and were both multidrug-resistant (MDR; i.e. resistant to at least rifampicin and isoniazid). We used PacBio and Illumina WGS to reconstruct the full circular genome and reconstitute the phylogeny of this novel lineage, which we named Lineage 8 (L8), and further investigate molecular and evolutionary events associated with the emergence of the MTBC.

Results

L8 related TB patient in Rwanda

The strain in Rwanda was isolated from a male patient, aged 77 years, HIV-negative, resident of Rulindo district bordering with the Southwest of Uganda, and who had lived in Uganda previously. The patient was diagnosed with rifampicin-resistant TB and the Xpert MTB/RIF assay (Xpert; Cepheid, Sunnyvale, CA, USA) showed a rare delayed probe B reaction (~3% prevalence in Rwanda)[14], later confirmed (see below) to be due to the Asp435Tyr mutation in the *rpoB* gene[15,16].

Per routine practice, the patient was initiated on standard short-course MDR-TB treatment[17]. However, the patient developed hypotension, and eventually died due to probable cardiac failure, after 20 days of treatment. Phenotypic drug-susceptibility testing (DST) confirmed resistance to both rifampicin and isoniazid, and susceptibility to other anti-TB drugs including ethambutol, fluoroquinolones, and second-line injectables.

Growth characteristics and biochemical properties of the Rwandan strain

The strain from Rwanda was grown in 12.5 days on Mycobacterial Growth Indicator Tubes. Colonies were observed on the fifth week after initial inoculation on Löwenstein-Jensen medium, indicating a slow grower phenotype with rough colonies (Figure 1). The strain also displayed archetypal biochemical characteristics of *M. tuberculosis sensu stricto*, including niacin production combined with urease hydrolysis (Table 1).

Genotypic resistance and SNP profile of the Rwandan strain by Deeplex Myc-TB

Following the MDR-TB diagnosis, the strain was included in the first set of tests for an ongoing MDR-TB diagnostic trial “DIAgnostics for MDR-TB in Africa (DIAMA) Clinicaltrials.gov, NCT03303963”, evaluating a new targeted deep sequencing assay, called Deeplex-MycTB

(GenoScreen, Lille, France). Deeplex-MycTB testing confirmed the presence of the *rpoB* Asp435Tyr mutation conferring rifampicin resistance, along with the *inhA* Ser94Ala mutation conferring isoniazid resistance, consistently with the MDR phenotype identified by phenotypic DST (**Figure 2**). This strain also harboured two alleles in phylogenetic positions in *embB* (Ala378) and *gidB* (Ala205) not associated with resistance to ethambutol or streptomycin, which were both shared by several MTBC lineages (L1, 5, 6, 7, and animal lineages) and *M. canettii*[18]. In addition, nine other - so far uncharacterized - SNPs were identified in six of the 18 gene targets interrogated by the assay (**Figure 2**). Moreover, this test detected an atypical spoligotype pattern, 1111100000000000000000000000000001110000000 (**Figure 2**), which was further confirmed by conventional membrane-based spoligotyping testing. This spoligotype pattern was unique in the global spoligotype database that comprises 111,637 MTBC isolates from 131 countries[19].

WGS analysis and phylogenetic position of the Rwandan and Ugandan strains

Results from WGS analysis of the Rwandan strain using Illumina sequencing confirmed all Deeplex-MycTB findings.

The strain isolated in Uganda was discovered independently upon screening global, publicly available genome datasets, where it was misclassified as *M. bovis* isolated from a human patient[20]. These WGS data revealed a similar spoligotype 1111100000000000000000000000000001111000111, characterized by the presence of spacers 1 to 5 and 34 to 37 (vs 34-36 in the Rwandan strain) with all intervening spacers missing. Moreover, the Ugandan strain also shared the same *rpoB* Asp435Tyr and *inhA* Ser94Ala mutations and the same sequence alleles in *embB* and *gidB.* The Ugandan strain contained an additional *katG* Ser315Thr mutation conferring isoniazid resistance, as well as the *embA* C-11A and *embB* Asp328Tyr mutations, and two *pncA* missense mutations, predictive of pyrazinamide resistance. Moreover, only three of the nine aforementioned uncharacterized SNPs detected by Deeplex-MycTB were shared between both strains.

To further assess the relationships between both strains and in comparison to other MTBC strains, a maximum likelihood phylogeny was inferred from 241 MTBC genomes, including representatives of all known human- and animal- adapted lineages[6] and using a *M. canettii* strain as an outgroup. This reconstruction revealed a unique phylogenetic position of the two new genomes from Rwanda and Uganda (**Figure 3**), representing a newly characterised monophyletic clade in which none of the known MTBC genomes are contained. Based on the phylogeny, this clade shares a MRCA with the rest of the MTBC, thus representing a new sister clade to the known MTBC, which we named Lineage 8 (L8). Comprehensive SNP analysis identified a total of 189 SNPs separating both genomes, which is within the range of zero to 700 SNPs found between any two strains within any of the lineages 1 to 7 of the MTBC[11]. The absence of any matching pattern in the global spoligotype database, as well as the lack of detection of this clade in previous large WGS datasets of MTBC strains from global sources, indicate that L8 is generally rare and geographically restricted to the African Great Lakes region. Specifically, the L8 spoligotype signature and the 3 SNPs specifically shared by both L8 strains were not detected in any of 115 MTBC genomes from a previous drug resistance survey in Uganda[21], nor in 380 rifampicin-resistant strains from Rwanda collected between 1991 and 2018, from routine drug resistance surveillance as well as various drug resistance surveys[22–24]. Furthermore, among 14 other isolates out of 27 from Uganda and Rwanda tested by Gene Xpert MTB/RIF that showed the same delayed probe B as L8, none displayed the L8 signatures when tested by Deeplex Myc-TB or by classical spoligotyping. Likewise, none of > 1,500 clinical samples from TB patients tested by Deeplex-MycTB from a recent nationwide drug resistance survey performed in Democratic Republic of Congo displayed the L8 spoligotype signature or the specific SNPs (data not shown).

Defining features of a complete L8 genome

To further assess the sister position of L8 and split from the remaining MTBC inferred from SNP analysis, the Rwandan strain was subjected to long read-based PacBio sequencing. Comparison of the obtained assembly with 36 available complete genomes of MTBC members, comprising L1-L4 (including H37Rv), *M. africanum* (L6) and *M. bovis* strains, showed a highly syntenic organization, with no major structural rearrangement between both groups. Although the assembled L8 genome of 4,379,493 bp was within the 4.34-4.43 Mb size range of the other MTBC genomes, it was 30 kb smaller than the 4.41-Mb mean size of genomes of *M. tuberculosis sensu stricto*[25]. However, the largest part of this gap was accounted by the absence of three genomic regions in L8, corresponding to regions of difference (RDs) known to be variably present or absent in other MTBC (sub)lineages[26,27](Supplementary Table 2). These include a 9.3-kb PhiRv1 prophage region (RD3), as well as 10.0-kb and 8.5-kb segments corresponding to RD14 and RD5, comprising the *plcABC* gene cluster and the *plcD* gene regions, respectively[26]. In L8, each of the two latter regions only contained one copy of the IS*6110* insertion sequence, devoid of direct repeats (DRs) that normally flank IS*6110* copies after transposition, indicating that these deletions in L8 resulted from recombination between two adjacent IS*6110* copies with loss of the intervening sequences[28]. These mobile DNA-related deletions, which also arose independently in several other MTBC branches[26,29], probably occurred after the divergence of L8 from the other MTBC lineages.

Conversely, a particular 4.4 kb genome region was present in both genomes of L8 and *M. canettii*, but absent in all other known members of the MTBC (Supplementary Table 3). This region comprises the *cobF* gene (**Figure 4**), encoding the precorrin 6A synthase involved in the cobalamin/vitamin B12 synthesis, along with two other genes, respectively encoding a PE-PGRS protein family member and a protein of unknown function. This region is present in the *M. canettii* genomes, as well as in the phylogenetically proximal non-tuberculous mycobacterial species *M. marinum* and *M. kansasii* (Supplementary Table 3). This ancestral locus was thus most likely lost in the MRCA of the other MTBC lineages, after its divergence from L8. However, none of the almost 900 other genes specifically identified in the *M. canettii* genomes, and absent in the other MTBC genomes, were found in the L8 genome, supporting the close relationship with the previously known MTBC branches indicated by the SNP-based phylogeny.

Further evidence for the early branching of L8 relative to the rest of the MTBC comes from examination of interrupted coding sequences (ICDSs), putatively reflecting molecular scars inherited during progressive pseudogenization of the MTBC genomes[30,31]. Four orthologues of MTBC ICDSs were previously found to be intact in the genomes of *M. canettii* strains, as well as in *M. marinum* and *M. kansasii* [4]. One of these four orthologues (*pks8*), which belongs to a multigene family encoding polyketide synthases involved in the biosynthesis of important cell envelope lipids[32], was also intact in the genomes of both L8 strains (Figure 5 and Supplementary Table 4). Moreover, we found an additional orthologue of MTBC ICDSs (i.e. *rv3899c-rv3900c*), coding for a conserved hypothetical protein, which was intact in the genomes of *M. canettii*, *M. kansasii*, *M. marinum* and both L8 strains (Supplementary Table 4). These two molecular scars were also likely acquired by the other MTBC lineages after their divergence from the common progenitor shared with L8.

The assembled L8 genome also included 48 out of 50 genes (the exception are rv3513c encoding the probable fatty-Acid-Coa ligase FadD18 and the PhiRv1 region; see above) present in MTBC members but not found in any of the STB genomes, including a number of genes putatively acquired through horizontal gene transfer by the common ancestor of the MTBC after its separation from *M. canettii*[4] (Supplementary Table 5). Likewise, consistent with the rough colony morphotype of the Rwandan strain, both L8 strains displayed the single polyketide-synthase-encoding *pks5* gene configuration shared by all MTBC members, instead of the dual *pks5* conformation found in *M. canettii* strains involved in the smooth colony phenotype of the latter strains[12]. Thus, the recombination between the two *pks5* genes and the loss of the intervening *pap* gene, thought to have resulted in surface remodelling and incremental gain of virulence after the phylogenetic separation from *M. canettii*[12], already existed in the common progenitor of L8 and the rest of the MTBC. Moreover, both L8 strains also contained the intact TbD1 and RD9 regions, shared by the other "ancestral" *M. tuberculosis* lineages (L1, L7) but subsequently lost by the so-called "modern" lineages of *M. tuberculosis* (TbD1 lost in L2-4), *M. africanum* (L5 and L6) and the animal lineages (RD9)[26].

In contrast to the highly clonal structure of the MTBC, *M. canettii* strains are highly recombinogenic, as apparent from mosaic sequence arrangements in their genomes and functional DNA transfer between *M. canettii* strains mediated by a distributive conjugal transfer (DCT)-like mechanism[4,33]. However, no significant genome-wide recombination signal was detected by ClonalFrameML analysis[34] between L8 and other MTBC strains (data not shown).

Discussion

The discovery of L8 provides unique insights into an ancestor of the MTBC that existed after the *pks5*-recombination-mediated surface remodelling, which occurred after separation of the MTBC MRCA from the *M. canettii* clade, but preceded the loss of the *cobF* region and gene interruptions in a later common ancestor of the other MTBC lineages. The seeming restriction of this lineage to the African Great Lakes region represents new evidence supporting an origin for the MTBC in the eastern part of the African continent. These findings reinforce results from previous work suggesting an East- rather than a West African origin of the MTBC[3,4,8,9,35].

A distinct ecological niche, linked to a potential environmental reservoir, has been hypothesized to explain the marked geographic restriction of *M. canettii* strains to the Horn of Africa, the lower persistence of these strains in infection models as well as their genome mosaicism implying multiple DNA recombination events within the *M. canettii* strain pool[4,10]. However, our results indicate L8 is as clonal as the rest of the MTBC[3,13,29,36]. Moreover, multi-drug resistance in both L8 isolates, and their detection in human patients in both cases (with reported absence of previous TB history for the Rwandan patient), suggests prolonged exposure to antibiotic treatment, and human-to-human transmission of a drug-resistant strain, rather than infection from a non-human source. While based on only two initial strains, these results are consistent with the presumed scenario of a human rather than a zoonotic origin for the TB disease[26,37].

The observation that both L8 strains share two uncommon rifampicin- and isoniazid-resistance conferring mutations in *rpoB* and *inhA* suggests that multidrug resistance was already acquired in their common ancestor. Isoniazid and rifampicin were introduced in TB treatments in Rwanda and Uganda in the late fifties and early nineties, respectively.

Therefore, the ~100 SNPs distance separating these two strains from their MRCA would imply a rapid molecular clock for L8, above the upper limit of 2.2 SNPs/genome/year most recently estimated for other MTBC clades[38]. However, this mutation rate cannot be confirmed until additional L8 samples are uncovered.

Remarkably, the absence of other L8 strains in datasets from Uganda, Rwanda and DRC, together comprising more than 2,000 strains, suggests that L8 is rare even within the African Great Lakes region. Such scarcity is compatible with selective sweeps of later branching MTBC strains, introduced more recently into the region. Similar scenarios have also been proposed to explain the slow apparent replacement of MTBC L5 and L6 by L4 in West Africa[39–41] and the restriction of L7 to Northern Ethiopia[42].

Loss-of-function linked to the deletion of *cobF* is a plausible candidate molecular event involved in such a replacement scenario for L8. Indeed, loss-of-function appears to be an important mechanism driving the pathoadaptive evolution of the TB pathogen, as shown for the role of the loss of lipo-oligosaccharide production (via recombination in the *pks5* locus)[12] in the evolution towards increased virulence from *M. canettii* to MTBC strains. Likewise, loss of secretion of PPE-MPTR and PE_PGRS proteins by the type VII secretion system ESX-5 (via mutations of the *ppe38* locus) has been involved in the hypervirulence of recent branches of L2 (alias "modern" Beijing) strains[43]. The loss of the *cobF* region in the other MTBC lineages, inferred from comparative genomics with *M. canettii* and non-tuberculous mycobacteria[4], was previously hypothesised to reflect enhanced adaptation to an intracellular parasitic lifestyle[44]. Indeed, the cobalamin/vitamin B12 synthesis pathway, of which the *cobF*-encoded precorring-6a synthase is a component, represents a highly complex and energy consuming process with about 30 enzymatic steps[45]. While the absence of this component may not entirely ablate cobalamin biosynthesis[46,47], its loss might have resulted in gain of fitness and reflect enhanced pathogenic professionalisation, by economical reliance upon the mammalian host environment as source of vitamin B12. If true, more recently emerged or introduced, *cobF*-deleted strains might conceivably have largely outcompeted L8 strains.

Our genomic data, on an as-yet-unknown ancestral stage between the MTBC and the putative progenitor pool of *M. canettii*-like mycobacteria, suggest further experiments to examine candidate molecular events potentially involved in the pathoadaptive evolution of *M. tuberculosis*. The discovery of such rare strains raises the possibility for the existence of further extant strains, especially in Eastern Africa, representing other clades further closing the biological gap between the MTBC and *M. canettii*.

Methods

Phenotypic characterization

We studied conventional mycobacterial growth and biochemical characteristics including colony morphology, niacin production, nitrate reduction, p-nitro benzoic acid growth inhibition, catalase production, urea hydrolysis, tween 80 hydrolysis, and thiophene carboxylic acid hydrazide growth inhibition[48]. For comparative purpose, a reference set of the seven known human-adapted MTBC lineages[49], together with *M. canettii (BCCM/ITM2018-C02321), M. bovis (BCCM/ITM960770), M. bovis BCG* (BCCM/ITMM006705), and *M. orygis* (BCCM/ITM2018-01492) strains were processed with the novel strain isolated in Rwanda. Moreover, phenotypic drug susceptibility testing to first- and second-line anti-TB drugs was done using the proportion method[50].

Targeted and whole genome sequencing

For targeted sequencing using the Deeplex-MycTB assay[16] and short-read Illumina-based WGS, a bead beating method was used to extract DNA from colonies (Supplementary method 1). Libraries of Deeplex-MycTB amplicons or genome fragments were constructed using the Nextera XT kit and sequenced on an Illumina MiSeq platform with paired end, 150-bp read lengths (Illumina, CA, USA). DNA extraction suitable for PacBio SMRT sequencing was performed using the Genomic DNA Buffer Set (Qiagen Inc, Germantown, Maryland, USA) (Supplementary method 2). Sequencing was performed on a PacBio RS II using the SMRT technology.

Deeplex-MycTB analysis and spoligotyping

Analysis of the Deeplex-MycTB sequencing data, including SNP calling and spoligotype identification, was performed by read mapping on *M. tuberculosis* H37Rv sequence references, using a parameterized web application (GenoScreen)[16]. Membrane-based spoligotyping was performed as described previously[51].

Illumina whole genome sequencing analysis

Raw genomic reads from the newly sequenced L8 genome from Rwanda and the L8 genome from Uganda (SAMN02567762) were processed as previously described[52]. Briefly, the reads were trimmed with Trimmomatic v0.33.22[53] and reads larger than 20 bp were kept. The software SeqPrep (https://github.com/jstjohn/SeqPrep) was used to identify and merge any overlapping paired-end reads. The resulting reads were aligned to the reconstructed ancestral sequence of the MTBC[54] using the mem algorithm of BWA v0.7.13[55]. Duplicated reads were marked using the MarkDuplicates module of Picard v2.9.1 (https://github.com/broadinstitute/picard) and local realignment of reads around InDels was performed using the RealignerTargetCreator and IndelRealigner modules of GATK v3.4.0[56]. SNPs were called with Samtools v1.2 mpileup[57] and VarScan v2.4.1[58] using the following thresholds: minimum mapping quality of 20, minimum base quality at a position of 20, minimum read depth at a position of 7X, maximum strand bias for a position 90%.

The spoligotype pattern of the strain from Uganda was extracted *in silico* from the raw reads using kvarQ[59].

Phylogenetic reconstruction

The maximum likelihood phylogeny was inferred with RAxML v.8.2.8[60] using an alignment containing only polymorphic sites and the branch lengths of the tree were rescaled using invariant sites (rescaled_branch_length = (branch_length * alignment_length) / (alignment_length +invariant_sites))[38,61].

A position was considered polymorphic if at least one genome had a SNP at that position. Deletions and positions not called according to the minimum threshold of 7x were encoded as gaps. We excluded positions with more than 20% missing data, positions falling in PE-PGRS genes, phages, insertion sequences and in regions with at least 50 bp identity to other regions in the genome. We also excluded variable positions falling in drug resistance-related genes. The phylogeny was computed using the general time-reversible model of sequence evolution (-m GTRCAT -V options), 100 bootstrap inferences and *M. canettii* (SRR011186) was used as an outgroup to root the phylogeny.

Whole genome *de novo* assembly, annotation and comparative genomics

Raw PacBio reads obtained from the Rwandan strain were assembled with Canu v1.6[62], using default settings and an expected genome size of 4.4 Mbp, typical of MTBC strains. After discarding 60,272 reads below minimal quality parameters, 106,681 reads were used for the assembly, resulting in mean coverage of 186x, 39x and 38x, after read correction, trimming, and unitigging, respectively. The obtained unique contig of 4,387,285 bp was circularized with Circlator v1.5.5[63] using default settings, resulting in an assembly of 4,379,493 bp. Additional sequence verification and correction was then performed by mapping Illumina reads obtained from the same strain, using pacbio-utils version 0.2[64] (https://github.com/douglasgscofield/PacBio-utilities) and snippy version 4.4[65] (https://github.com/tseemann/snippy). Alignments of the final assembly were performed against an ensemble of complete genome sequences available from 38 strains of tubercle bacilli. This set included 34 *M. tuberculosis* strains from lineages 1, 2, 3 and 4 (comprising H37Rv), *M. africanum* L6 GM041182, *M. bovis* AF2122/97, as well as the closest STB-A (CIPT 140010059) and most distant (STB-K) *M. canettii* strains (Supplementary Table 1). Comparative alignments and genome annotation were performed based on BLAST searches and analysis of gene synteny, using Artemis and Artemis comparison tool[66], as well as a custom Multiple Annotation of Genomes and Differential Analysis (MAGDA) software previously used for annotation of *M. canettii* and *Helicobacter pylori* genomes[4,67]. Comparisons with orthologues from *M. canettii* STB-D, -E, -G, -H, -I, and -J in addition to STB-A and -K, and from *M. marinum* type strain M and *M. kansasii* genomes were additionally done using the Microscope platform v3.13.3[68]. When applicable, annotations were transferred from those of *M. tuberculosis* or *M. canettii* orthologs in the TubercuList/Mycobrowser database, using BLAST matches of > 90% protein sequence identity, an alignable region of >80% of the shortest protein length in pairwise comparisons and visual inspection of the gene synteny. ACT comparison files were generated using MAUVE 2015-02-25 software to visualize the genome-wide distribution of SNP densities between the assembled L9 genome from Rwanda and *M. tuberculosis* H37Rv and *M. canettii* STB-A and STB-K genomes. Recombination between L8 and other MTBC lineages or *M. canettii* was assessed from a progressive MAUVE alignment of the PacBio assembled L8 genome and previously published closed genomes[65] using ClonalFrameML[34].

Accession codes

The complete genome sequence of the L8 strain from Rwanda was deposited in the NCBI repository under project PRJNA598991 with SRR10828835 and SRR10828834 accession codes for Illumina- and PacBio-derived genome sequences, respectively.

References

1. Gagneux, S. Ecology and evolution of Mycobacterium tuberculosis. *Nature Reviews Microbiology* **16**, 202–213 (2018).

2. World Health Organization. *Global Tuberculosis Report 2018*. (2018).

3. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nature genetics* **45**, 1176–82 (2013).

4. Supply, P. *et al.* Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. *Nature genetics* **45**, 172–9 (2013).

5. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–7 (2014).

6. Brites, D. *et al.* A New Phylogenetic Framework for the Animal-Adapted Mycobacterium tuberculosis Complex. *Frontiers in Microbiology* **9**, 2820 (2018).

7. Gagneux, S. *et al.* Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 2869–73 (2006).

8. Wirth, T. *et al.* Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS pathogens* **4**, e1000160 (2008).

9. Gutierrez, M. C. *et al.* Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. *PLoS pathogens* **1**, e5 (2005).

10. Blouin, Y. *et al.* Progenitor "Mycobacterium canettii" clone responsible for lymph node tuberculosis epidemic, Djibouti. *Emerging infectious diseases* **20**, 21–8 (2014).

11. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in Mycobacterium

tuberculosis. *Seminars in Immunology* **26**, 431–444 (2014).

12. Boritsch, E. C. *et al.* pks5-recombination-mediated surface remodelling in Mycobacterium tuberculosis emergence. *Nature Microbiology* **1**, 15019 (2016).

13. Boritsch, E. C. *et al.* Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 9876–81 (2016).

14. Ng, K. C. S. *et al.* Automated algorithm for early identification of rifampicin-resistant tuberculosis transmission hotspots in Rwanda [abstract]. *The International Journal of Tuberculosis and Lung Disease* **22**, 605 (2018).

15. Tagliani, E. *et al.* Culture and Next-generation sequencing-based drug susceptibility testing unveil high levels of drug-resistant-TB in Djibouti: results from the first national survey. *Scientific reports* **7**, 17672 (2017).

16. Makhado, N. A. *et al.* Outbreak of multidrug-resistant tuberculosis in South Africa undetected by WHO-endorsed commercial tests: an observational study. *The Lancet Infectious Diseases* **18**, 1350–1359 (2018).

17. Trébucq, A. *et al.* Treatment outcome with a short multidrug-resistant tuberculosis regimen in nine African countries. *The International Journal of Tuberculosis and Lung Disease* **22**, 17–25 (2018).

18. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nature Communications* **5**, (2014).

19. Couvin, D., David, A., Zozio, T. & Rastogi, N. Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the Mycobacterium tuberculosis genotyping database. *Infection, Genetics and Evolution* (2018). doi:10.1016/j.meegid.2018.12.030

20. Wanzala, S. I. *et al.* Retrospective Analysis of Archived Pyrazinamide Resistant

Mycobacterium tuberculosis Complex Isolates from Uganda-Evidence of Interspecies Transmission. *Microorganisms* **7**, (2019).

21. Ssengooba, W. *et al.* Whole genome sequencing to complement tuberculosis drug resistance surveys in Uganda. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **40**, 8–16 (2016).

22. Carpels, G. *et al.* Drug resistant tuberculosis in sub-Saharan Africa: an estimation of incidence and cost for the year 2000. *Tubercle and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* **76**, 480–6 (1995).

23. Umubyeyi, A. N. *et al.* Results of a national survey on drug resistance among pulmonary tuberculosis patients in Rwanda. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* **11**, 189–94 (2007).

24. World Health Organization & WHO. WHO | Global tuberculosis report 2016. *WHO* (2019).

25. Yang, T. *et al.* Pan-genomic study of Mycobacterium tuberculosis reflecting the primary/secondary genes, generality/individuality, and the interconversion through copy number variations. *Frontiers in Microbiology* **9**, (2018).

26. Brosch, R. *et al.* A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 3684–9 (2002).

27. Tsolaki, A. G. *et al.* Functional and evolutionary genomics of Mycobacterium tuberculosis: Insights from genomic deletions in 100 strains. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4865–4870

(2004).

28. Brosch, R. *et al.* Genomic analysis reveals variation between Mycobacterium tuberculosis H37Rv and the attenuated M. tuberculosis H37Ra strain. *Infection and Immunity* **67**, 5768–5774 (1999).

29. Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W. & Small, P. M. Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4871–6 (2004).

30. Deshayes, C. *et al.* Detecting the molecular scars of evolution in the Mycobacterium tuberculosis complex by analyzing interrupted coding sequences. *BMC evolutionary biology* **8**, 78 (2008).

31. Smith, N. H., Hewinson, R. G., Kremer, K., Brosch, R. & Gordon, S. V. Myths and misconceptions: the origin and evolution of Mycobacterium tuberculosis. *Nature Reviews Microbiology* **7**, 537–544 (2009).

32. Etienne, G. *et al.* Identification of the polyketide synthase involved in the biosynthesis of the surface-exposed lipooligosaccharides in mycobacteria. *Journal of Bacteriology* **191**, 2613–2621 (2009).

33. Boritsch, E. C. & Brosch, R. Evolution of Mycobacterium tuberculosis: New Insights into Pathogenicity and Drug Resistance. in *Tuberculosis and the Tubercle Bacillus, Second Edition* **4**, 495–515 (American Society of Microbiology, 2016).

34. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS computational biology* **11**, e1004041 (2015).

35. Hershberg, R. *et al.* High Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. *PLoS Biology* **6**, e311 (2008).

36. Supply, P. *et al.* Linkage disequilibrium between minisatellite loci supports clonal

evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area. *Molecular Microbiology* **47**, 529–538 (2003).

37. Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A. & Behr, M. A. Genomic Deletions Suggest a Phylogeny for the *Mycobacterium tuberculosis* Complex. *The Journal of Infectious Diseases* **186**, 74–80 (2002).

38. Menardo, F., Duchêne, S., Brites, D. & Gagneux, S. The molecular clock of Mycobacterium tuberculosis. *PLoS pathogens* **15**, e1008067 (2019).

39. Niobe-Eyangoh, S. N. *et al.* Genetic biodiversity of Mycobacterium tuberculosis complex strains from patients with pulmonary tuberculosis in Cameroon. *Journal of Clinical Microbiology* **41**, 2547–2553 (2003).

40. Godreuil, S. *et al.* First molecular epidemiology study of Mycobacterium tuberculosis in Burkina Faso. *Journal of Clinical Microbiology* **45**, 921–927 (2007).

41. Groenheit, R. *et al.* The Guinea-Bissau Family of Mycobacterium tuberculosis Complex Revisited. *PLoS ONE* **6**, e18601 (2011).

42. Comas, I. *et al.* Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa. *Current Biology* **25**, 3260–3266 (2015).

43. Ates, L. S. *et al.* Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. *Nature Microbiology* **3**, 181–188 (2018).

44. Boritsch, E. C. *et al.* A glimpse into the past and predictions for the future: the molecular evolution of the tuberculosis agent. *Molecular Microbiology* **93**, 835–852 (2014).

45. Martens, J. H., Barg, H., Warren, M. & Jahn, D. Microbial production of vitamin B12. *Applied Microbiology and Biotechnology* **58**, 275–285 (2002).

46. Gopinath, K. *et al.* A vitamin B $_{12}$ transporter in *Mycobacterium tuberculosis*. *Open*

Biology **3**, 120175 (2013).

47. Minias, A., Minias, P., Czubat, B. & Dziadek, J. Purifying Selective Pressure Suggests the Functionality of a Vitamin B12 Biosynthesis Pathway in a Global Population of Mycobacterium tuberculosis. *Genome Biology and Evolution* **10**, 2326–2337 (2018).

48. LEÃO, S. C. *et al.* Practical handbook for the phenotypic and genotypic identification of mycobacteria. (2004).

49. Borrell, S. *et al.* Reference set of Mycobacterium tuberculosis clinical strains: A tool for research and product development. *PLOS ONE* **14**, e0214088 (2019).

50. Kent, P. & Kubica, G. *Public Health Mycobacteriology: A Guide for the Level III Laboratory US Department of Health and Human Services, Centres for Disease Control*. (1985).

51. van der Zanden, A. G. M. *et al.* Improvement of differentiation and interpretability of spoligotyping for Mycobacterium tuberculosis complex isolates by introduction of new spacer oligonucleotides. *Journal of clinical microbiology* **40**, 4628–39 (2002).

52. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018).

53. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**, 2114–20 (2014).

54. Comas, I. *et al.* Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nature genetics* **42**, 498–503 (2010).

55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–60 (2009).

56. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for

analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297–303 (2010).

57. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* **27**, 2987–93 (2011).

58. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568–76 (2012).

59. Steiner, A., Stucki, D., Coscolla, M., Borrell, S. & Gagneux, S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics* **15**, 881 (2014).

60. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

61. Duchene, S. *et al.* Inferring demographic parameters in bacterial genomic data using Bayesian and hybrid phylogenetic methods. *BMC evolutionary biology* **18**, 95 (2018).

62. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* **27**, 722–736 (2017).

63. Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology* **16**, 294 (2015).

64. Scofield, D. G. GitHub - douglasgscofield/PacBio-utilities: Collection of utilities for working with PacBio-based assemblies. Available at: https://github.com/douglasgscofield/PacBio-utilities. (Accessed: 20th December 2019)

65. Seemann, T. GitHub - tseemann-snippy Rapid bacterial SNP calling and core genome alignments. Available at: https://github.com/tseemann/snippy.

(Accessed: 20th December 2019)

66. Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics (Oxford, England)* **24**, 2672–6 (2008).

67. Li, H. *et al.* East-Asian Helicobacter pylori strains synthesize heptan-deficient lipopolysaccharide. *PLoS Genetics* **15**, (2019).

68. Médigue, C. *et al.* MicroScope-an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data. *Briefings in bioinformatics* **20**, 1071–1084 (2019).

Acknowledgements

Calculations were partially performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel. This work was supported by EDCTP2 grant DRIA2014-326—DIAMA of the European Union, the Belgian General Directorate for Development Cooperation (PhD fellowship to JCSN), grant ANR-16-CE35-0009 from Agence Nationale de la Recherche, the Swiss National Science Foundation (grants 310030_188888, IZRJZ3_164171, IZLSZ3_170834 and CRSII5_177163), and the European Research Council (309540-EVODRTB).

The views and opinions of authors expressed herein do not necessarily state or reflect those of EDCTP. P.S., C.G. and S.D. declare the following competing interests: P.S. was a consultant of Genoscreen; C.G. and S.D. were employees of the same company. The other authors declare no competing interests. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.G., P.S., C.M., B.C.d.J., L.R. and J.C.S.N. designed the study. P.S., J.C.S.N., C.L., M.M., C.M. and S.G. analyzed data and wrote the manuscript, with comments from all authors. A.J. and C.M. performed the assembly of sequences. M.M. annotated the L9 genome. C.L., F.M., D.B. and A.J. performed SNP analyses and phylogenetic reconstruction. M.M. and P.S. conducted comparative analyses of complete mycobacterial genomes, with support from C.L. and O.T. J.C.S.N., E.B.N., I.M.H., J.B.M., W.M., K.F. and M.D. performed and/or analyzed data from mycobacterial isolation, growth assays, phenotypic characterization and/or molecular tests. S.D., C.G., J.C.S.N., E.B.N., E.A. and M.K.K. conducted targeted deep sequencing analyses. S.D., C.G. and W.S. L. Majlessi, F.S., C. Locht and C. Leclerc conducted and/or analyzed immune assays. J.T., A. Criscuolo and S.B. conducted MLST, recombination and/or phylogenetic analyses. L.F. conducted histopathological analyses. V.K., M.O. and C.P. created bioinformatics tools and analyzed data. M.F. isolated STB strains. T.S. and T.P.S. conducted core genome and NeighborNet analyses.

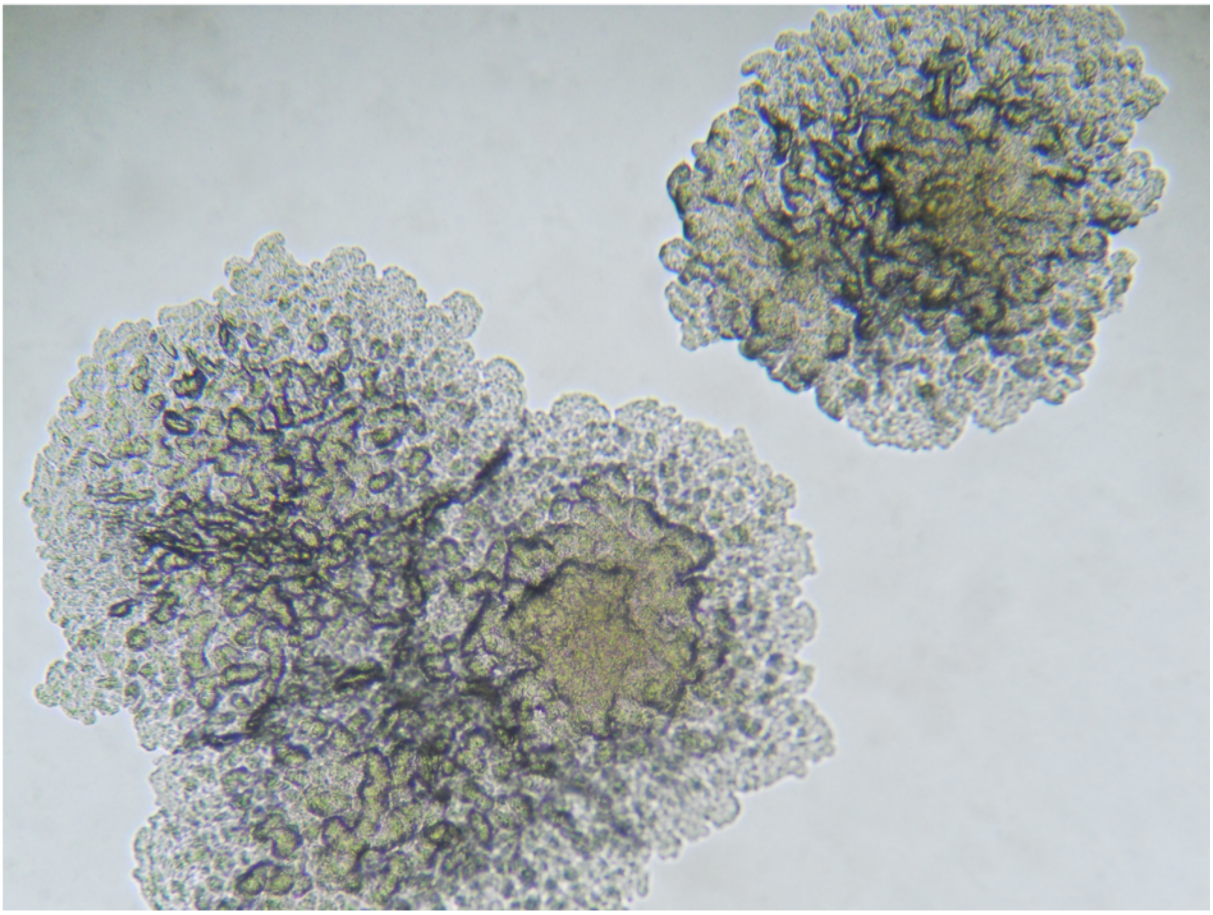

Figure 1. Microscopic image of L9 on Dubos agar medium showing typical rough colonies (read at 100x).

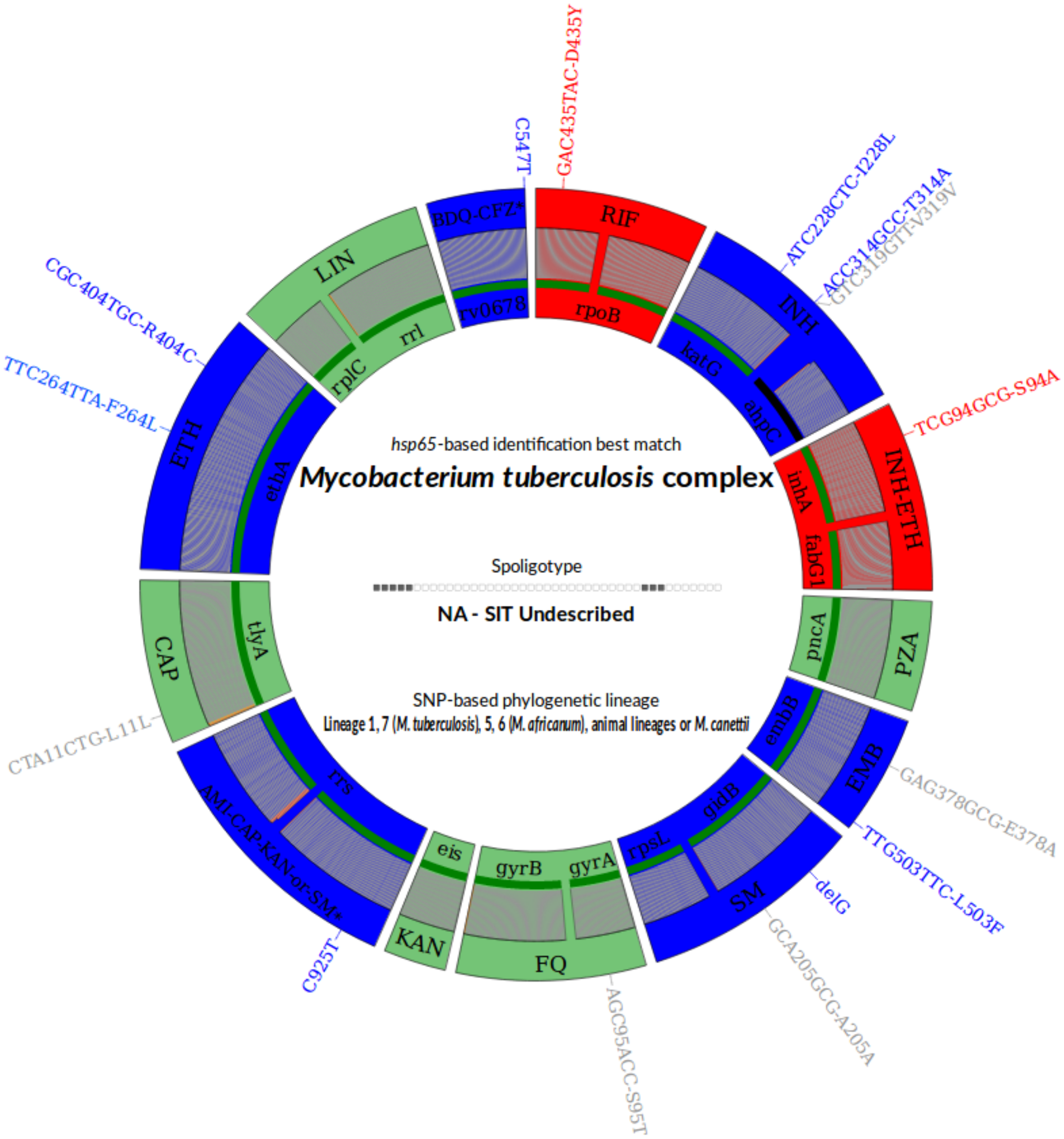

Figure 2. Deeplex-MycTB results identifying a multidrug resistant strain from Rwanda with an atypical genotypic profile in the *M. tuberculosis* complex. Target gene regions are grouped within sectors in a circular map according to the tuberculous drug resistance with which they are associated. The two sectors in red indicate regions where rifampicin and isoniazid resistance associated mutations are detected. The multiple sectors in blue refer to regions where as yet uncharacterized mutations are detected, while sectors in green indicate regions where no mutation or only mutations not associated with resistance (shown in gray around the map) were detected. Green lines above gene names represent the reference sequences with coverage breadth above 95%. Limits of detection

(LOD) of potential heteroresistance (reflected by subpopulations of reads bearing a mutation), depending on the coverage depths over individual sequence positions, are indicated by grey (LOD 3%) and orange zones (variable LOD >3%–80%) above the reference sequences. Information on an unrecognized spoligotype, an equivocal SNP-based and on mycobacterial species identification, based on *hsp65* sequence best match, are shown in the centre of the circle. *AMI, amikacin; BDQ, bedaquiline; CAP, capreomycin; CFZ, clofazimine; EMB, ethambutol; ETH, ethionamide; FQ, fluoroquinolones; KAN, kanamycin; LIN, linezolid; INH, isoniazid; PZA, pyrazinamide; RIF, rifampin; SM, streptomycin; SIT, spoligotype international type.

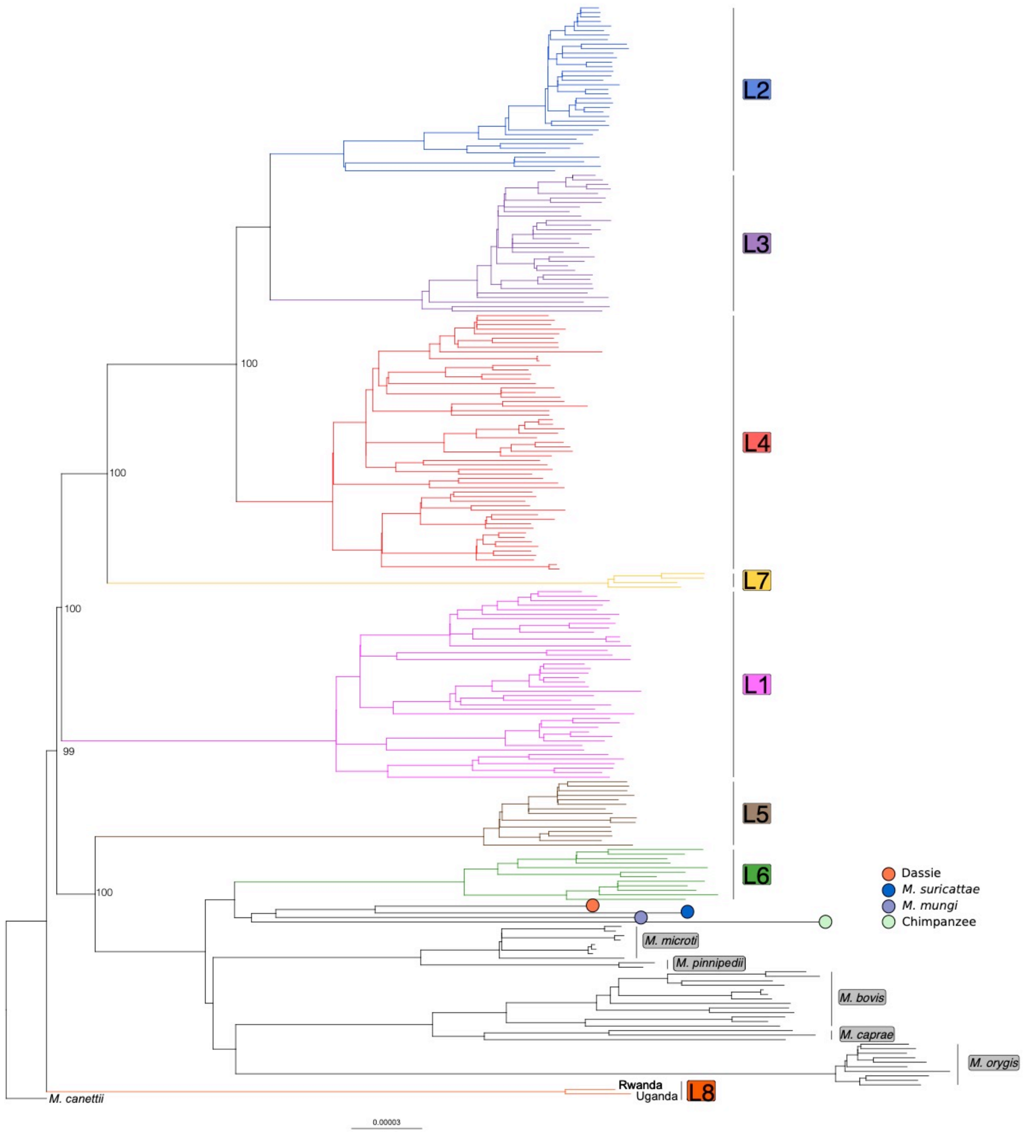

Figure 3. Maximum likelihood phylogeny of 241 MTBC genomes, inferred from 43,442 variable positions. The scale bar indicates the number of substitutions per polymorphic site. Branches corresponding to human-adapted strains are coloured and branches corresponding to animal-adapted strains are depicted in black. The phylogeny is rooted on *M. canettii* and bootstrap values are shown for the most important splits.

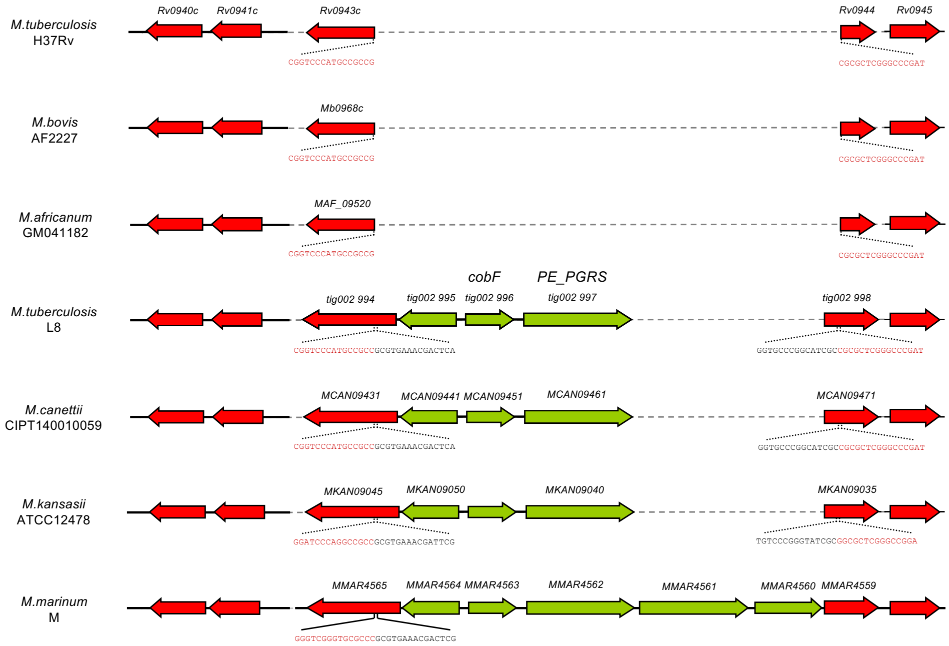

Figure 4. Aligned genome segments showing the *cobF* gene region in *M. tuberculosis* L8, M. canettii CIPT140010059 (alias STB-A), M. kansasii ATCC12478 and M. marinum M strains, and the corresponding deletion in *M. tuberculosis* H37Rv, M. bovis AF2122/97, and *M. africanum* GM041182. Coding sequences of this region are shown in green, and flanking coding sequences in red. Sequences flanking the deletion point in truncated genes in *M. tuberculosis*, *M. africanum* and *M. bovis*, and in the cobF region present in *M. canettii*, *M. kansasii* and *M. marinum* are indicated in red and black, respectively. Dashed lines correspond to missing segment parts relatively to the longest segment found in *M. marinum*.

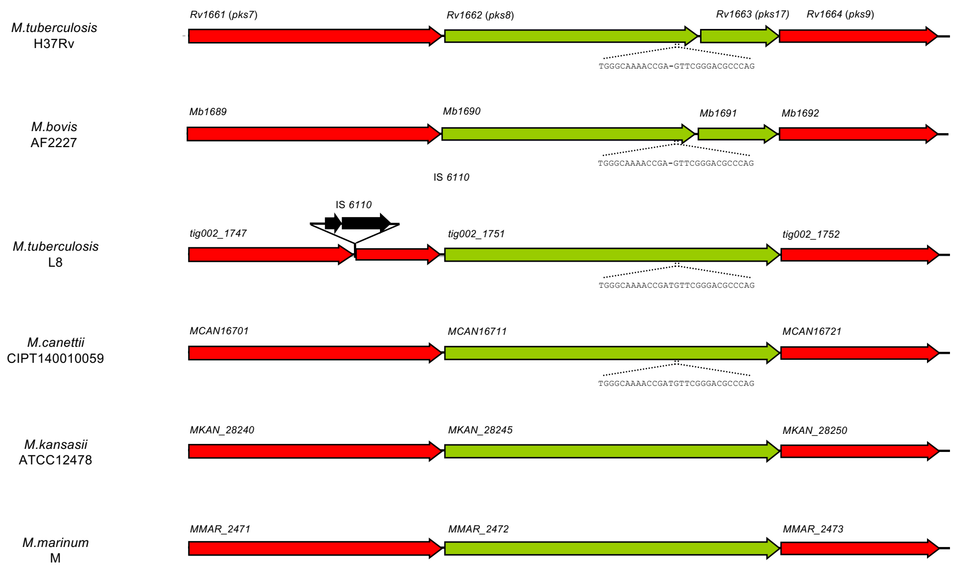

Figure 5. Aligned genome segments showing the interrupted coding sequences *pks8/17* in *M. tuberculosis* H37Rv and *M. bovis* AF2122/97 gene region, and complete *pks8* genes in L8, *M. canettii* CIPT140010059 (alias STB-A) and *M. kansasii* ATCC12478. *pks8/17* and *pks8* coding sequences are shown in green, and flanking genes in red. Sequences flanking the 1-nucleotide deletion and resulting in a frameshift in *M. tuberculosis* complex strains are indicated. Dashed lines correspond to missing segment parts relatively to the longest segment found in *M. canettii.*

Table 1. Standard biochemical characteristics of selected mycobacterial species or *M. tuberculosis* complex lineages/subspecies versus L8

Mycobacterial species/lineage	BCCM/ITM REF N^o	Niacin production	Nitrate reduction	Urease hydrolysis	Tween hydrolysis	Catalase production	Arylsulfatase
L8	2018-01172	+	+	+	weak +	-	-
L1	2018-00082	+	+	+	-	weak +	-
L2	2018-00087	+	-	+	-	-	-
L3	2018-00089	+	+	+	-	-	-
L4	2018-00093	+	+	+	+	-	-
L5	2018-00095	+	-	-	+	-	-
L6	2018-00099	+	-	weak +	weak +	-	-
L7	2018-00101	+	-	weak +	-	-	-
M. bovis	960770	-	-	weak +	-	-	-
M. bovis BCG	M006705	-	-	+	-	weak +	-
M. orygis	2018-01492	+	-	+	-	-	-
M. canettii	C02321	-	-	+	+	weak +	-
M. fortuitum	C1	-	+	+	-	weak +	+

+: positive reaction; -: negative