S1 File. Supplementary Methods

Customized script and markers used for identification of MTBC species and lineages

Species and lineage classification were confirmed using RDs and specific markers according to previously published data [3, 4]. For MTBC species, we used a blastn search of primer sets against each genome and product sizes previously described by Warren et al [5]. For *M. tuberculosis* and *M. africanum* lineage identification, we used primer sets and product sizes compiled by Gagneux et al [6] from previous work [7–9]. When identifying species, the following RDs were expected to be absent in *M. tuberculosis*: RD1, RD4, RD9, and RD12, while in *M. africanum*, RD1, RD4, RD12 and **RD1<sup>nc</sup>** should be absent and RD9 should be present [5]. For *M. tuberculosis* L1, L2, L3 and L4 the following markers were expected to be present: RD239, RD105, RD750, or **pks15/1 7 bp deletion**, respectively [7–9]. For *M. tuberculosis* L7, TbD1 and RD239 were expected to be absent [7–9]. And finally, the presence of RD711 and RD702 corresponded to *M. africanum* L5 and L6, respectively [7–9].

For *M. bovis*, the absence of RD1 and presence of RD4, RD9 and RD12 were used for species identification [5]. *Mycobacterium bovis* lineage/clonal complex classification of each genome was retrieved from previously published data [3, 4], which are based on the following markers: presence of **RDAf2** for clonal complex African 2 (Af2) [10], SNP in the *guaA* gene (at 3,765,573 position in the *M. bovis* AF2122/97 reference strain) and deletion of spoligotype spacer 21 for clonal complex European 2 (Eu2) [11], and presence of **RDEu1** (RD17) for clonal complex European 1 (Eu1) [12]. These markers correspond to the newly defined lineages Lb1, Lb3 and Lb4 of *M. bovis*, respectively [3].

Exclusion of genomes based on the sequencing platform used

All complete genomes of *M. tuberculosis*, and draft and complete genomes of *M. africanum*, *M. bovis* and *M. canettii* available in RefSeq as of January 2019 were screened. Due to potential sequencing errors arising from homopolymeric tracts [1, 2], we did not include genomes that were sequenced with 454, Ion Torrent or PacBio platforms, even if they were sequenced and assembled in a hybrid form with Illumina. We also did not include genomes that were sequenced many years ago with Sanger only, such as the *M. tuberculosis* CDC1551 (NC_002755.2). There were only two exceptions that were kept: the *M. tuberculosis* H37Rv strain
(NC_018143.2), which was sequenced in a hybrid form (Illumina, 454 and Sanger) and *M. canettii* CIPT140010059 (NC_015848.1), which was sequenced with Sanger and Illumina. These genomes were kept because they were both complete and reference genomes of the species and were also not outliers in the number of pseudogenes. We also excluded two *M. tuberculosis* genomes from Narayann et al. 2013 (strains EAI5/NITR206, Beijing/NITR203) because they had extremely high number of pseudogenes, and because two other genomes included in the same publication did not meet quality standards to be included in the RefSeq database. This curation based on sequencing platforms and sequencing quality resulted in the exclusion of 43 genomes (28 *M. tuberculosis*, 1 *M. africanum*, 10 *M. bovis* and 4 *M. canettii*) from an initial dataset of 201 that met all other inclusion criteria set for this study and described in the main text of the manuscript. Thus, 158 genomes of the MTBC and *M. canettii* were used in this study.

**Identification, retrieval, and criteria for inclusion of pseudogenes**

Pseudogene DNA sequence retrieval was performed using a customized Python script (Pseudo_retriev.py) that takes as input the *gbff* (Genbank Genome File) and *cds fna* (Fasta Nucleic Acid) files (Fig. 1A and B). The *gbff* file contains the genome annotation and features, DNA sequence and the positional coordinates of genes and pseudogenes. Coordinates of the pseudogenes reported in the *gbff* file cover the maximal extent of the aligned frameshifted protein. The corresponding *locus_tag* of each pseudogene described in the *gbff* file was then used to retrieve their DNA sequences from the *fna* file to create a pseudogene dataset of the studied genomes.
Fig. 1. Overview of the method used to analyze pseudogenes of the *Mycobacterium tuberculosis* complex (MTBC) and *Mycobacterium canettii*. (A) Types and number of dataset files of MTBC and *M. canettii* genomes retrieved from NCBI (National Center for Biotechnology Information). Genome files of *Mycobacterium tuberculosis* (n=85), *Mycobacterium africanum* (n=27), *Mycobacterium bovis* (n=41) and *M. canettii* (n=5) were downloaded from RefSeq database and processed in a pipeline for pseudogene retrieval. (B) Pseudogene curation pipeline. Paralogous pseudogenes and those identified as belonging to the PE/PPE family, mobile genetic elements (insertion sequences, transposases, phage, integrases, and maturases) were removed. Then, the corresponding DNA sequence dataset of pseudogenes that had full-length gene counterparts of each strain was generated and analyzed for the pseudogenization rate, events that led to pseudogenization and conservation of pseudogenes among strain (i.e. non-redundancy).

The NCBI’s annotation tool, PGAP, reports interrupted genes at the end of contigs of draft genomes as pseudogenes. For being assembly artefacts, these pseudogenes were then removed from our datasets. Briefly, the annotation of such pseudogenes, which contains the following note: “too short partial abutting assembly gap”, was used in our algorithm (Pseudo_retriev.py) to exclude these pseudogenes from the analyses during pseudogene retrieval (Fig. 1A and B). A correlation graph between the number of contigs and pseudogenes before and after this exclusion was generated using Seaborn library of Python [13] (Fig. 2). From the 25,837 pseudogenes, 4,793 pseudogenes fitted this criterium in draft genomes of *M. bovis*, *M. africanum*, and *M. canettii* and were removed (Fig. 3). Before and after comparisons indicate that the number of pseudogenes from draft genomes increases with the number of contigs, showing a positive correlation error, which was normalized after the elimination of these “false” pseudogenes (Fig. 2).
The number of contigs was correlated with number of pseudogenes before and after removal of pseudogenes at contigs ends. A total of 4,793 of 25,837 pseudogenes were excluded. (A) *Mycobacterium africanum*, (B) *Mycobacterium bovis*, (C) *Mycobacterium canetti*. Orange: before removing “false pseudogenes” predicted at the end of contigs. Blue: after removing the “false pseudogenes” predicted at the end of contigs.

In the next step of the curation of the dataset, we removed pseudogenes identified as belonging to the PE/PPE family and mobile genetic elements (insertion sequences, transposases, phage, integrases, and maturases) and paralogous pseudogenes. This was accomplished by removing any pseudogene that was PGAP-annotated according to a list of names (Table 1) and by selecting any pseudogene for which a full-length gene counterpart (explained below) was a PE/PPE gene located in problematic repetitive areas identified by Marin et al. [14] (PE/PPE DNA seq file, Fig. 3). This resulted in the exclusion of 5,653 pseudogenes from the dataset of 21,044 redundant pseudogenes (Fig. 3). We then used CD-HIT-est [15] to cluster DNA sequences of the pseudogenes to find paralogous pseudogenes. The parameters used were 90% identity and 80% coverage of the longest gene in that cluster. A total of 8 pseudogenes had N’s and 563 formed paralogous groups, which were removed from the analysis; 14,820 pseudogenes remained in the dataset (Fig. 3).
| Annotations | PPE family protein PPE28 |
|-------------|--------------------------|
| DDE transposase | PPE family protein PPE3 |
| HNH endonuclease | PPE family protein PPE32 |
| insertion sequence B9 | PPE family protein PPE33 |
| Insertion sequence IS21 ATP-binding protein | PPE family protein PPE34 |
| integrase | PPE family protein PPE37 |
| IS110 family transposase | PPE family protein PPE38 |
| IS110 family transposase IS1547 | PPE family protein PPE42 |
| IS110-like element IS1547 family transposase | PPE family protein PPE44 |
| IS1380 family transposase | PPE family protein PPE46 |
| IS1634 family transposase | PPE family protein PPE49 |
| IS1634 family transposase ISMyca1 | PPE family protein PPE50 |
| IS200/IS605 family transposase | PPE family protein PPE54 |
| IS21 family transposase | PPE family protein PPE57 |
| IS21 family transposase ISMbo1 | PPE family protein PPE58 |
| IS21-like element ISM22 family transposase | PPE family protein PPE59 |
| IS256 family transposase | PPE family protein PPE6 |
| IS256 family transposase IS1081 | PPE family protein PPE65 |
| IS256 family transposase IS1553 | PPE family protein PPE66 |
| IS256 family transposase IS1554 | PPE family protein PPE67 |
| IS256-like element IS1081 family transposase | PPE family protein PPE69 |
| IS256-like element IS1553 family transposase | PPE family protein PPE7 |
| IS256-like element IS1554 family transposase | PPE family protein PPE8 |
| IS3 family transposase | PE family protein PPE21 |
| IS3-like element IS987 family transposase | PE family protein PPE24 prohead protease |
| IS3 family transposase | type VII secretion system ESX-1 target PE35 |
| IS3-like element IS987 family transposase | type VII secretion system ESX-1 target PPE68 |
| IS3 family transposase | type VII secretion system ESX-3 target PE5 |
| IS4 family transposase | type VII secretion system ESX-3 target PPE4 |
| IS4 family transposase | type VII secretion system ESX-5 target PE18 |
| IS5 family transposase | type VII secretion system ESX-5 target PPE25 |
| IS5/IS1182 family transposase | type VII secretion system ESX-5 target PPE26 |
| IS5/IS1182 family transposase | type VII secretion system ESX-5 target PPE27 |
| IS607 family transposase | mycofactocin radical SAM maturase |
| IS607 family transposase | group II intron reverse transcriptase/maturase |
| IS607 family transposase | site-specific integrase |
| IS607 family transposase | phage Gp37/Gp68 family protein |
| IS607 family transposase | phage holin family protein |
| IS607 family transposase | phage major capsid protein |
| IS607 family transposase | phage portal protein |
| IS607 family transposase | phage terminase |
| IS607 family transposase | phage terminase small subunit P27 family |
| ISNCY family transposase | bacteriophage protein |
| PE domain-containing protein | HK97 family phage |
| PE family protein | PPE domain-containing protein |
| PE family protein PE1 | PPE family protein |
| PE family protein PE27 | PPE family protein PPE1 |
| PE family protein PE3 | PPE family protein PPE10 |
| PE family protein PE4 | PPE family protein PPE12 |
| PE PGRS family protein | PPE family protein PPE13 |
| PE-PGRS family protein | PPE family protein PPE17 |
| PE-PPE domain-containing protein | PPE family protein PPE18 |
| phage integrase | PPE family protein PPE19 |
| PPE family protein PPE21 | PPE family protein PPE20 |
| PPE family protein PPE22 | PPE family protein PPE24 prohead protease |
To posterior analyses, we only considered pseudogenes that had a full-length gene counterpart (except for those pseudogenes present in ≥90% of the strains). For this, we run CDHIT-est again (with the same parameters) with the 14,820 pseudogenes and all genes from the studied strains; the aim was to eliminate pseudogene redundancy and detect full-length gene counterparts. For those pseudogene clusters in which a full-length gene counterpart was not found using this approach, we also used blastn of these pseudogenes against the database of genes using 90% identity and 80% query coverage parameters (this approach improved the detection of full-length gene counterparts for some of the incomplete genes). A total of 1,740 clusters containing pseudogenes were detected. Using both approaches, full-length gene counterparts were not detected in 861 clusters (Fig. 3). These were removed from the dataset. Thus, the remaining 879 non-redundant pseudogenes (corresponding to 10,262 redundant pseudogenes) were considered in all analyses of pseudogenes in this study (Fig. 3). Nineteen of these 879 non-redundant pseudogenes were conserved in more than 90% of the strains used in this study, 510 were present in less than 90% of the strains but in at least two strains and 369 were singletons (i.e. present in only one strain).
**Fig. 3.** Flowchart of the pseudogene retrieval and curation to build the final dataset of analyzed pseudogenes of *Mycobacterium tuberculosis* complex strains and *Mycobacterium canettii*. *equivalent to 10,262 redundant pseudogenes.*

**Position of pseudogenes in contigs of draft genomes**

To further evaluate if the predicted pseudogenes were randomly distributed along the length of contigs of draft genomes, we gathered the locus_tag of each pseudogene and retrieved (from the *gbff* file) the size of the contig in which this pseudogene was located and its position inside the contig. Contigs were divided into four equal parts from 5' to 3', irrespective of its size in bp, and the number of pseudogenes in each part was computed. The number of detected pseudogenes according to size of contigs in draft genomes was also calculated. Fig. 4 shows that the pseudogenes are randomly distributed along the length of contigs, without any biases towards their ends (Fig. 4A, B, C). In addition, only a minority of pseudogenes are located within contigs of up to 5,000 bp in *M. africanum* and *M. bovis*. For all three ecotypes with draft genomes, the majority of pseudogenes are in large contigs (> 20,000 bp or 50,000 bp) (Fig. 4D,E,F).
Fig. 4. Localization of pseudogenes in contigs of draft genomes and by contig size. (A-C) Frequency distribution of pseudogenes along the length of contigs. Each contig was divided into four parts, from 5' to 3', irrespective of its size in bp. 0-5 represents the first quarter of the contig, 0.5-1 the second quarter, 1-1.5 the third quarter and 1.5-2 the last quarter of the contig. (D-E) Frequency distribution of pseudogenes according to contig size. (A and D) *Mycobacterium africanum*, (B and E) *Mycobacterium bovis*, (E and F) *Mycobacterium canettii*.

Quality control of the number of CDS in draft genomes

The pseudogenization rate calculated in this study takes into consideration the number of CDS in its denominator. As to understand if there is any influence in the number of predicted CDS and the number of contigs, we generated correlation plots for *M. africanum*, *M. bovis* and *M. canettii* genomes (Fig. 5). We found no correlation between the number of contigs and the number of CDSs (Fig. 5). In addition, MTBC genomes used in this study have overlapping number of CDS irrespective of their completeness [number of CDS in *M. tuberculosis* (all complete genomes): 3,959 and 4,143, *M. africanum* (draft genomes): 4,009 – 4,065, *M. bovis* (draft genomes): 3,750 – 4,187]. Thus, the influence of the number of CDS in the pseudogenization rate calculated in this study is expected to be negligible, particularly when also considering the high conservancy of the clonal MTBC genomes, which vary little in length.
Fig. 5. Correlation plots between the number of contigs and predicted coding DNA sequences (CDS) of draft genomes used in this study. (A) Mycobacterium africanum, (B) Mycobacterium bovis, (C) Mycobacterium canettii. Orange: number of CDS/genome before adding the number of truncated genes by the assembly gaps. Blue: number of CDS/genomes after adding the number of truncated genes by the assembly gaps.

Criteria for the selection of the BER (blast extend repraze) best hit

As to identify the best hit in the resulting BER alignments for each pseudogene, first, all BER hits that were shorter than the query length were excluded. Then, the % coverage for each BER hit was calculated as:

\[
\% \text{HitCoverage} = \left( \frac{\text{hit}_{\text{end-coordinate}} - \text{hit}_{\text{start-coordinate}}}{\text{hit}_{\text{pseudogen length}}} + 1 \right) \times 100
\]  

(1)

where \( \text{hit}_{\text{end-coordinate}} \) is the hit coordinate where the pseudogene ends, \( \text{hit}_{\text{start-coordinate}} \) is the hit coordinate where the pseudogene starts, and \( \text{hit}_{\text{pseudogen length}} \) is the pseudogene length (bp). The protein sequence of the best BER hit with the highest % coverage was used to replace the corresponding amino acid sequence of each pseudogene, creating a dataset of best BER hits of pseudogenes for each genome.

Code availability

All customized codes used in this study can be found at the GitHub repository: https://github.com/LaPAM-USP/Soler-Camargo-2022.

References

1. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 2012;30:434–439.
2. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, et al. A tale of three NGS
sequencing platforms. *BMC Genomics* 2012;13:13.

3. **Zimpel CK, Patané JSL, Guedes ACP, de Souza RF, Silva-Pereira TT, et al.** Global Distribution and Evolution of *Mycobacterium bovis* Lineages. *Front Microbiol* 2020;11:843.

4. **Zimpel CK, Brandão PE, de Souza Filho AF, de Souza RF, Ikuta CY, et al.** Complete Genome Sequencing of *Mycobacterium bovis* SP38 and Comparative Genomics of *Mycobacterium bovis* and *M. tuberculosis* Strains. *Front Microbiol* 2017;8:2389.

5. **Warren RM, Pittius NCG Van, Barnard M, Hesseling A, Engelke E, et al.** Differentiation of *Mycobacterium tuberculosis* complex by PCR amplification of genomic regions of difference. *Int J Tuberc Lung Dis* 2006;10:818–822.

6. **Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, et al.** Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* 2006;103:2869–73.

7. **Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, et al.** Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains. *Proc Natl Acad Sci U S A* 2004;101:4865–4870.

8. **Mostowy S, Onipede A, Gagneux S, Niemann S, Kremer K, et al.** Genomic analysis distinguishes *Mycobacterium africanum*. *J Clin Microbiol* 2004;42:3594–3599.

9. **Marmiesse M, Brodin P, Buchrieser C, Gutierrez C, Simoes N, et al.** Macro-array and bioinformatic analyses reveal mycobacterial ‘core’ genes, variation in the ESAT-6 gene family and new phylogenetic markers for the *Mycobacterium tuberculosis* complex. *Microbiology* 2004;150:483–496.

10. **Berg S, Garcia-Pelayo MC, Müller B, Hailu E, Asiimwe B, et al.** African 2, a Clonal Complex of *Mycobacterium bovis* Epidemiologically Important in East Africa. *J Bacteriol* 2011;193:670–678.

11. **Rodriguez-Campos S, Schürch AC, Dale J, Lohan AJ, Cunha M V., et al.** European 2 — A clonal complex of Mycobacterium bovis dominant in the Iberian Peninsula. *Infect Genet Evol* 2012;12:866–872.

12. **Smith NH, Berg S, Dale J, Allen A, Rodriguez S, et al.** European 1: A globally important clonal complex of *Mycobacterium bovis*. *Infect Genet Evol* 2011;11:1340–1351.

13. **Waskom M, Botvinnik O, O’Kane D, Hobson P, Lukauskas S, et al.** mwaskom/seaborn: v0.8.1 (September 2017). Epub ahead of print September 2017. DOI: 10.5281/ZENODO.883859.

14. **Marin M, Vargas R, Harris M, Jeffrey B, Epperson LE, et al.** Benchmarking the empirical accuracy of short-read sequencing across the *M. tuberculosis* genome. *Bioinformatics* 2022;38:1781–1787.

15. **Fu L, Niu B, Zhu Z, Wu S, Li W.** CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28:3150–3152.