MICROBIAL GENOMICS

RESEARCH ARTICLE

Soler-Camargo *et al.*, *Microbial Genomics* 2022;8:000876

DOI 10.1099/mgen.0.000876

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

The rate and role of pseudogenes of the *Mycobacterium tuberculosis* complex

Naila Cristina Soler-Camargo[1,2], Taiana Tainá Silva-Pereira[1], Cristina Kraemer Zimpel[1,2], Maurício F. Camacho[3], André Zelanis[3], Alexandre H. Aono[4,5], José Salvatore Patané[6], Andrea Pires dos Santos[7] and Ana Marcia Sá Guimarães[1,7,*]

Abstract

Whole-genome sequence analyses have significantly contributed to the understanding of virulence and evolution of the *Mycobacterium tuberculosis* complex (MTBC), the causative pathogens of tuberculosis. Most MTBC evolutionary studies are focused on single nucleotide polymorphisms and deletions, but rare studies have evaluated gene content, whereas none has comprehensively evaluated pseudogenes. Accordingly, we describe an extensive study focused on quantifying and predicting possible functions of MTBC and *Mycobacterium canettii* pseudogenes. Using NCBI's PGAP-detected pseudogenes, we analysed 25837 pseudogenes from 158 MTBC and *M. canetii* strains and combined transcriptomics and proteomics of *M. tuberculosis* H37Rv to gain insights about pseudogenes' expression. Our results indicate significant variability concerning rate and conservancy of *in silico* predicted pseudogenes among different ecotypes and lineages of tuberculous mycobacteria and pseudogenization of important virulence factors and genes of the metabolism and antimicrobial resistance/tolerance. We show that *in silico* predicted pseudogenes contribute considerably to MTBC genetic diversity at the population level. Moreover, the transcription machinery of *M. tuberculosis* can fully transcribe most pseudogenes, indicating intact promoters and recent pseudogene evolutionary emergence. Proteomics of *M. tuberculosis* and close evaluation of mutational lesions driving pseudogenization suggest that few *in silico* predicted pseudogenes are likely capable of neofunctionalization, nonsense mutation reversal, or phase variation, contradicting the classical definition of pseudogenes. Such findings indicate that genome annotation should be accompanied by proteomics and protein function assays to improve its accuracy. While indels and insertion sequences are the main drivers of the observed mutational lesions in these species, population bottlenecks and genetic drift are likely the evolutionary processes acting on pseudogenes' emergence over time. Our findings unveil a new perspective on MTBC's evolution and genetic diversity.

DATA SUMMARY

The authors confirm that all supporting data, code, and protocols have been provided in the article or supplementary data files. All customized codes used in this study are available at the GitHub repository: https://github.com/LaPAM-USP/Soler-Camargo-2022.

Received 04 February 2022; Accepted 13 July 2022; Published 17 October 2022

Author affiliations: [1]Laboratory of Applied Research in Mycobacteria, Department of Microbiology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, SP, Brazil; [2]Department of Preventive Veterinary Medicine and Animal Health, College of Veterinary Medicine, University of São Paulo, São Paulo, SP, Brazil; [3]Functional Proteomics Laboratory, Federal University of São Paulo (UNIFESP), São José dos Campos, SP, Brazil; [4]Center of Molecular Biology and Genetic Engineering, University of Campinas, Campinas, SP, Brazil; [5]Institute of Science and Technology, Federal University of São Paulo (UNIFESP), São José dos Campos, SP, Brazil; [6]Laboratory of Cellular Cycle, Butantan Institute, São Paulo, SP, Brazil; [7]Department of Comparative Pathobiology, College of Veterinary Medicine, Purdue University.

***Correspondence:** Ana Marcia Sá Guimarães, anamarcia@usp.br

Keywords: comparative genomics; *Mycobacterium tuberculosis* complex; pseudogenes; loss of function mutations; frameshift; phase variation.

Abbreviations: ASC, Ascertainment Bias Correction; BER, Blast Extend Repraze; BIC, Bayesian Information Criterion; CDS, Coding DNA Sequence; COG, Cluster of Orthologous Group; ENA, European Nucleotide Archive; FNA, Fasta Nucleic Acid; GBFF, GenBank Genome File; GFF, Gene Finding Format; GPL, Glycopeptidolipids; HGT, Horizontal Gene Transfer; IS, Insertion Sequence; Maf, Mycobacterium africanum; Mbo, Mycobacterium bovis; Mca, Mycobacterium canettii; ML, Maximum Likelihood; MS/MS, Tandem Mass Spectrometry; Mtb, Mycobacterium tuberculosis; MTBC, Mycobacterium tuberculosis complex; NCBI, National Center for Biotechnology Information; OADC, Oleic Acid, Albumin, Dextrose, and Catalase; PDIM, Phthiocerol Dimycoserate; PGAP, Prokaryotic Genome Annotation Pipeline; PGL, Phenolic Glycolipid; RD, Regions of Difference; RefSeq, Reference Sequence Database; SNP, Single Nucleotide Polymorphism; TB, Tuberculosis; VF, Virulence Factor; VFDB, Virulence Factor Database.

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

Impact Statement

Pseudogenization is one way in which bacteria can modulate gene content, possibly phenotype, but this process has never been comprehensively evaluated in the phenotypically distinct pathogens causing tuberculosis in humans and animals. We show that pseudogenes are a source of genetic variability to the MTBC, their rate varies according to ecotypes and lineages, and they are only moderately conserved among strains. Specific types of mutations were identified as drivers of pseudogenization, underscoring the importance of insertions, deletions, and mobile elements to the genetics of mycobacteria. Surprisingly, gene loci under pseudogenization correspond to ~16% of the global gene pool of MTBC. While pseudogenization may inactivate major virulence factors and other genes of the metabolism, sequences predicted as pseudogenes *in silico* may also reverse to its original gene sequence, serve as source of protein neofunctionalization, or carry a mutation with no significant impact on protein function, which contradict the classical definition of pseudogenes and warrant further studies. Corroborating previous knowledge about MTBC evolution, population bottlenecks and genetic drift are likely the acting forces that allow pseudogenes' emergence. Our study fills a gap about the MTBC evolution, providing the basis to further understand how these bacteria display different phenotypes of virulence and host tropism.

INTRODUCTION

Tuberculosis (TB) is one of the most devastating infectious diseases of humans and animals. It is caused by the *Mycobacterium tuberculosis* complex (MTBC), a bacterial group composed of 11 species or ecotypes. The MTBC can be divided into human (*M. tuberculosis*, *Mycobacterium africanum*) and animal-adapted pathogens (*Mycobacterium bovis*, *Mycobacterium caprae*, *Mycobacterium orygis*, *Mycobacterium microti*, *Mycobacterium mungi*, *Mycobacterium suricattae*, *Mycobacterium pinnipedii*, Chimpanzee bacillus, and Dassie Bacillus) [1]. The human-adapted strains of the MTBC are further classified into seven lineages; *M. tuberculosis* (L1-L4 and L7) and *M. africanum* (L5 and L6), based on phylogenomic analyses. During the development of this study, two other lineages were described: *M. tuberculosis* L8 and *M. africanum* L9 [2, 3]. Noteworthy, different lineages have distinct geographical spread, virulence, transmission capacity, and propensity to acquire drug resistance [4, 5]. For instance, *M. africanum* L5 and L6 are restricted to West Africa [6] and have lower disease progression rates than *M. tuberculosis* [7]. Animal-adapted pathogens infect a broad range of host species and establish animal reservoirs [8], with *M. bovis* and *M. caprae*, possibly *M. orygis*, regarded as the agents of 'zoonotic TB' [9–12]. In contrast, *M. tuberculosis* is highly adapted to humans [8]. Reasons for discrepancies underlying host tropism and virulence are elusive but likely a combination of host and bacterial factors.

The MTBC evolves under clonal evolution, and their genomes have high nucleotide identity (>99.95%) over homologous regions. Unlike non-tuberculous mycobacteria, the MTBC is not subjected to horizontal gene transfer (HGT) or significant recombination events [13, 14]. In the absence of HGT, members of the MTBC shape their gene content mainly through gene loss and duplication events [15], with a tendency for genomic decay [16–20]. Thus, despite low genetic variability, these microorganisms evolve by large deletions (some referred to as regions of difference, 'RD'), single nucleotide polymorphisms (SNPs), short indels (insertions or deletions), duplication of a limited number of paralogous gene families, and transposition of insertion sequence (IS) elements [21–23]. These alterations, albeit seemly small, translate into an array of virulence and host tropism phenotypes [8]. However, how exactly these mutations lead to alterations in the MTBC's adaptive traits remains unclear.

Most MTBC evolution studies focus on phylogenetic signatures given by detecting SNPs and RDs, while its gene repertoire has started to be analysed only recently [24–31]. When bacteria lose genes due to mutations, few disabled gene segments may remain; these non-functional segments are known as pseudogenes [32, 33]. Using assembled genomes, the detection of bacterial pseudogenes is based on *in silico* identification of genes containing frameshifts or premature stop codons, mutations frequently associated with gene inactivation [32, 34, 35]. This method has successfully identified pseudogenes in bacteria that were later experimentally confirmed and shown to provide adaptive traits to these bacteria to survive in new environmental niches [36, 37]. Hence, the extent and consequence of predicted pseudogenization should be explored to understand the genetic processes governing the maintenance or emergence of specific traits. MTBC genetic variability through comprehensive pseudogene analysis and characterization has not yet been explored. Therefore, this study aimed to quantify and functionally characterize *in silico* predicted pseudogenes among members of the MTBC. As an evolutionary comparative, genomes of *M. canettii* were also included.

METHODS

Dataset of MTBC and *M. canettii* genomes

All complete genomes of the MTBC (except for BCGs) and *M. canettii* deposited in RefSeq (Reference Sequence Database), NCBI (National Center for Biotechnology Information), as of January 2019, were selected, totalizing 117 genomes. Among these, the only ecotype with more than three complete genomes was *M. tuberculosis*. To increase the number of genomes from other MTBC ecotypes, both complete and draft genomes of *M. africanum* and *M. bovis* were included while keeping only complete

genomes of *M. tuberculosis*. Additionally, the nine available assembled genomes of *M. canettii* were selected to compare with a phylogenetically close bacterial species. The remaining MTBC ecotypes had less than three genomes each (complete or drafts) available at the time of the study, precluding statistically meaningful analyses, and thus, were not included in the study.

Quality inclusion criteria for a genome consisted of (i) availability in RefSeq database (to ensure the highest standard of sequencing and assembly possible from a public database), (ii) not sequenced with 454, Ion Torrent, PacBio or only-Sanger technologies, (iii) confirmation of species based on MTBC's RDs and phylogenetic analysis (described below), (iv) annotation with the latest version of the PGAP (Prokaryotic Genome Annotation Pipeline) of NCBI, and (v) more than 95% coverage compared to 'Bacteria dataset' in BUSCO (v.4.1.2) [38]. Genomes of 85 *M. tuberculosis* (all complete genomes), 27 *M. africanum* (27 draft genomes), 41 *M. bovis* (two complete genomes and 39 draft genomes), and five *M. canettii* (one complete genome and eight draft genomes) met inclusion criteria for this study (Table S1 and Supplementary Methods, available in the online version of this article). The number of contigs of MTBC and *M. canettii* draft genomes ranged from five to 191 and 502 to 587, respectively. The N50 of MTBC draft genomes varied from 44571 to 1456982, with a median of 122015 (Table S1).

Lineage classification and phylogenetic analysis

MTBC species and lineages (*M. tuberculosis* L1-L4 and L7, *M. africanum* L5 and L6, *M. bovis* Lb1-Lb4) were confirmed by searching for the RDs or SNP markers in the assembled genomes using blastn and a previously described Python script [39], respectively (Supplementary Methods). Additionally, a phylogenetic tree was built to confirm MTBC lineages further. A core-SNP matrix was generated using kSNP3 [40] and subjected to ascertainment bias correction (ASC) using the '-fconst' directive of IQ-Tree [41] as described [39]. The Modelfinder programme was used to select the best substitution model for the ASC-corrected SNP alignment according to Bayesian Information Criterion (BIC). The best-chosen model, HKY+I+G, was then fixed for maximum likelihood (ML) phylogenetic reconstruction using 1000 UFBoot pseudoreplicates [42]. Graphical customization of the phylogenetic tree was performed using Iroki [43]. *Mycobacterium canettii* CIPT140010059 was used as the outgroup. Lineage information from metadata associated with each BioSample was also used for confirmation when available.

Retrieval and criteria used to select pseudogenes

Pseudogenes reported by PGAP were used in this study. PGAP detects and annotates genes and pseudogenes using algorithms that combine *ab initio* gene prediction using GeneMarkS+ [44] and homology-based methods. The global alignment algorithm ProSplign [45] aligns predicted protein sequences with a reference genome for frameshift identification and delivers to GeneMarkS+. GeneMarkS+ [44] integrates the information about protein alignment, frameshifted genes, non-coding RNA, and typical DNA statistical patterns for protein and non-coding regions in gene prediction to report pseudogenes [46]. Once a pseudogene is identified, the original gene model is replaced with a new gene feature containing a 'pseudo' qualifier (Fig. S1).

DNA sequences of the pseudogenes were retrieved using a customized Python script (Pseudo_retriev.py) that takes as input the *gbff* (Genbank Genome File) and *cds fna* (Fasta Nucleic Acid) files of the selected genomes (Supplementary Methods). Genes located at the 5′ and 3′ ends of contigs of draft genomes may be falsely reported by PGAP as pseudogenes due to truncations/interruptions caused by the assembly gap. Thus, they were excluded from the initial dataset (Pseudo_retriev.py; Supplementary Methods) and their quantity added to the total number of CDS (coding DNA sequences) in each genome.

Posterior analysis included only non-paralogous pseudogenes that had a full-length gene counterpart in at least one other strain (except for those present in ≥90% of the strains) (Supplementary Methods). Pseudogenes identified as mobile genetic elements (e.g. transposases, integrases, insertion sequences - IS) or from the PE/PPE gene family were also excluded from the dataset (Supplementary methods).

Pseudogenization rate and events that led to pseudogenization

All pseudogenes that met the criteria mentioned above were counted for each genome in their corresponding dataset of pseudogene DNA sequences (Pseudo_retriev.py; Supplementary Methods). The rate of pseudogenization (%) was calculated as:

$$\%genomepseud = \frac{N_{pseud}}{N_{CDS}} \times 100$$

where $^{N}{}_{pseud}$ is the number of pseudogenes and $^{N}{}_{cds}$ is the number of CDS in each genome (corrected by the addition of missing CDS at the end of contigs described above).

Pseudogenization rates were further evaluated by ancestral character estimation onto the generated ML tree using R software with the function ace() of the package phytools [47]. Tree mapping and plotting of the trait's evolution were built using the contMap() and plot() functions, respectively [47]. Optimization was done fixing *M. canettii* as the outgroup.

Three major types of events leading to pseudogenization are annotated by PGAP: frameshift, incomplete gene (a gene disruption caused by an insertion sequence (IS) or transposition process), and internal stop codon (characterized by a premature truncation due to a nonsense mutation). The number of events of each genome was retrieved and counted from the *gbff* file and compared

among bacterial species, according to the number of pseudogenes that met inclusion criteria. As pseudogenes may be annotated as having more than one event, a non-exclusive quantification was performed.

Heatmap of conservation of pseudogenes

A heatmap of pseudogenes presence and absence in each genome was constructed from a non-redundant dataset of pseudogenes that met inclusion criteria. Briefly, pseudogene redundancy was eliminated by clustering pseudogene DNA sequences using CD-HIT-est [48] with parameters of 90% identity and 80% coverage of the longest gene in that cluster (Supplementary Methods). Results were organized into a matrix of presence or absence of each non-redundant pseudogene according to the strains in columns. This matrix was then transformed into a heatmap using the Seaborn library [49]. Corresponding Rv numbers (*M. tuberculosis* H37Rv, NC_00962.3, AL123456.3) for each of the non-redundant pseudogenes were recovered using CD-HIT-est or blastn against the gene sequences of *M. tuberculosis* H37Rv using the same parameters described above.

Estimation of pan-genome

The pan-genome of the bacterial strains used in this study was estimated with Orthofinder v2.3.8 [50] using the *cds faa* files containing the whole proteome (without pseudogenes) from each strain. The total number of detected orthologous groups of core and accessory genomes and unique genes of each strain were counted to compose the pan-genome.

Protein functions affected by pseudogenization

Due to the degenerated nature of predicted amino acid sequences of pseudogenes, full-length protein counterparts of each pseudogene were used to infer their functional annotation. Briefly, the nucleotide sequence of each pseudogene was analysed using the BER (Blast Extend Repraze) algorithm provided in Manatee software [51], which applies a blastx search against the UniRef100 database. The resulting pairwise alignments served as input to a modified Smith-Waterman algorithm that aligns an extended translated nucleotide sequence of the query (300 nucleotides downstream and upstream of each gene sequence) against the protein hits of the blastx search. This alignment allows frameshift detection and corrects the reading frame according to the reference protein, providing a superior blast hit than a regular blastp search for pseudogenes. The BER tool also extends the alignment through these potential frameshifts or in-frame stop codons, providing a more robust homology-based similarity search. The best hit in the resulting BER alignments for each pseudogene was selected for further analysis (see selection criteria in Supplementary Methods).

The datasets of best BER hits of pseudogenes and the whole proteome of each strain were analysed with EggNOG [52] to provide COG (clusters of orthologous group) functional classification [53] in a non-exclusive way. The proportions of genes and pseudogenes in each COG category were compared using Pearson's Chi-squared Test by enrichment analysis of COG categories, with correction for multiple comparisons, and the null hypothesis was rejected at 5% for all species. Differences were considered statistically significant when $P \leq 0.05$ and one degree of freedom.

Pseudogenes of each strain with a BER hit successfully annotated with EggNOG were displayed on a heatmap by COG classification. Briefly, a new matrix of non-redundant pseudogenes was generated as described above and adding the COG classification of each pseudogene's best BER hit. The corresponding functional classification was performed in a non-exclusive way. The heatmap was built using the Seaborn library [49].

Corresponding Rv numbers with known functional annotation (i.e. gene names) and present in ≥2 genomes were used in STRING database [54] to detect functional protein associated networks among the non-redundant pseudogenes.

Gene essentiality

Gene essentiality data of *M. tuberculosis* [55] and *M. bovis* [56] were used to evaluate if the detected pseudogenes are considered essential for growth *in vitro*. Corresponding Rvs of the non-redundant pseudogenes were searched against these datasets and reported as essential, or providing growth advantage or disadvantage, irrespective of the ecotype in which they were detected.

Pseudogenization of virulence factors

A list of protein sequences (n=227) related to virulence factors (VFs) in MTBC and *M. canettii* was downloaded from the Virulence Factor Database (VFDB) [57] and used as a reference dataset for the identification of pseudogenized VF in the 158 bacterial strains. VF amino acid sequences were searched against pseudogene DNA dataset using tblastn with ≥95% identity and ≥80% query coverage parameters. The presence and absence of pseudogenized VFs were displayed in a heatmap generated using the Seaborn library [49].

RNA-Seq data processing

We selected a previously published [58] RNA-Seq dataset obtained from *M. tuberculosis* H37Rv RNA extracted from mid-logarithmic phase growth in Middlebrook 7H9-OADC (oleic acid, albumin, dextrose, and catalase) broth. RNA-Seq reads

(PRJEB23469) from six replicates were retrieved from the European Nucleotide Archive (ENA) (R1 to R6: ERR2299755, ERR2299756, ERR2299757, ERR2299758, ERR2299759, ERR2299760). HISAT2 [59] (Galaxy Version 2.1.0+galaxy5 [60]) was used to map the reads against *M. tuberculosis* H37Rv reference genome (NC_018143.2) using default parameters, which include exclusion of reads mapping in more than one genomic region. As pseudogenes do not have regular gene *IDs*, a *gff* (Gene Finding Format) file containing pseudogene coordinates and sequential ids (from the *cds fna* file) was constructed. The six BAM files (from read mapping) and the pseudogene *gff* file were used as input to HTSeq [61] (Galaxy Version 0.9.1 [60]). Read counting (i.e. read depth) was performed using the htseq_count function according to default parameters and a Phred 20 quality scale.

Proteomics data processing

We selected previously published [62] combined proteomic data obtained from *M. tuberculosis* H37Rv extracted from 5 and 6 weeks growth in Middlebrook 7H9-OADC and Proskauer-Beck broth, respectively. Proteomic data were accessed via ProteomeXchange Consortium/PRIDE partner repository [63] with the dataset identifier PXD010956. Mass spectrometric (RAW) data were analysed within the Trans Proteomics Pipeline platform [64] (v.5.2.0; Build 201903130949–7900). Briefly, RAW files were converted to the mzXML file format and searched with Comet [65] (version 2018.01, rev. 4) and X!Tandem search engines [66] against an in-house database containing the multi-fasta protein sequences of *M. tuberculosis* H37Rv (NC_018143.2; a total of 4235 entries from whole proteome), the predicted pseudogenes, and 57 full-length gene counterparts of pseudogenes (i.e. pseudogene amino acid sequences with the corrected frame). Peptide identification was based on a search with mass deviation of the precursor ion of 20 ppm and the fragment mass tolerance was set to 0.2 Da. Enzyme specificity was set to trypsin and at least two missed cleavages were allowed. Cysteine carbamidomethylation was selected as fixed modification whereas methionine oxidation, glutamine/asparagine deamidation, acetylated N-termini, and methionine formylation were selected as variable modifications. Protein identification was accepted after estimating the False Discovery Rate (FDR) calculated based on the score distributions in the output of the Comet search engine for each biological replicate. Search results were combined using iProphet [67] tool and further filtered with PeptideProphet and ProteinProphet to a >99% confidence interval, corresponding to a FDR of less than 1%. Protein identifications were accepted if they contained at least one identified unique peptide.

Statistical analyses

The number of pseudogenes, pseudogenization rate, and pseudogenization events were compared among groups using the non-parametric Kruskal-Wallis test, followed by the Dunn test to detect differences between two groups, in GraphPad Prism 6 [68]. Differences were considered statistically significant when $P \leq 0.05$. Boxplots were generated using Matplotlib and Seaborn libraries of Python [49].

RESULTS

Pseudogenization rates vary between and within-species of the MTBC and *M. canettii*

A total of 10262 pseudogenes meeting inclusion criteria of the 158 MTBC and *M. canettii* strains were analysed in this study (Table S1 and Supplementary Methods). Most pseudogenes (68.62%) have less than 1086 bp, while the remaining 31.38% have between 1086 and 12640 bp (Fig. S2). When considering draft genomes, pseudogenes are randomly distributed along the length of contigs, without any biases towards their ends, and mostly present in contigs >20000 or 50000 bp (Supplementary Methods). Genomes of *M. tuberculosis* showed lower number of pseudogenes/genome and median pseudogenization rate compared to other species ($P \leq 0.001$), while no statistical difference was observed between *M. africanum*, *M. bovis* and *M. canettii* strains ($P > 0.05$) (Fig. 1a and c). These results indicate that pseudogenization rates vary between species of the MTBC.

We then investigated if within-species variations were associated with specific MTBC lineages. Out of the 85 *M. tuberculosis* genomes, one (1.18%) was identified as L1, 10 (11.76%) as L2, and 74 (87.06%) as L4 (Fig. 2). Genome representatives of *M. tuberculosis* L3 and L7 were not identified in the dataset. The outlier among *M. tuberculosis* genomes was the L1 strain (Fig. 1c). Only two (7.41%) out of the 27 genomes of *M. africanum* were identified as L5, while the remaining 25 (92.59%) were L6. Using a novel lineage classification of *M. bovis* [39], of the 41 genomes, one strain (2.44%) was classified as Lb1, five (12.19%) as Lb3, 31 (75.61%) as Lb4, and four (9.76%) as unknown (Fig. 1a and c). Potential segregation of the number of pseudogenes was observed between *M. tuberculosis* lineages, but not for *M. africanum* or *M. bovis* lineages (Fig. 1c). The pseudogenization rate of *M. tuberculosis* L2 was significantly higher than L4 ($P \leq 0.001$) (Fig. 1b), suggesting that L2, and possibly L1 strains, contribute to increase the pseudogenization in *M. tuberculosis* at the population level. Ancestral character estimation of the pseudogenization rate on the ML tree indicated a tendency for less pseudogenization in the ingroup (MTBC) compared to the outgroup (*M. canettii*), with the weakest pseudogenization being towards the origin of *M. tuberculosis* L4 (Fig. S3 and S4).

Frameshift mutations are the main cause of pseudogene formation in the MTBC

Frameshift, incomplete gene, and internal stop codon events occurred in 66.52%, 32.17%, and 16.12% of the 10262 pseudogenes, respectively (Table S1). Multiple events were observed in only 10.23% of the pseudogenes. Frameshifts were the main driver

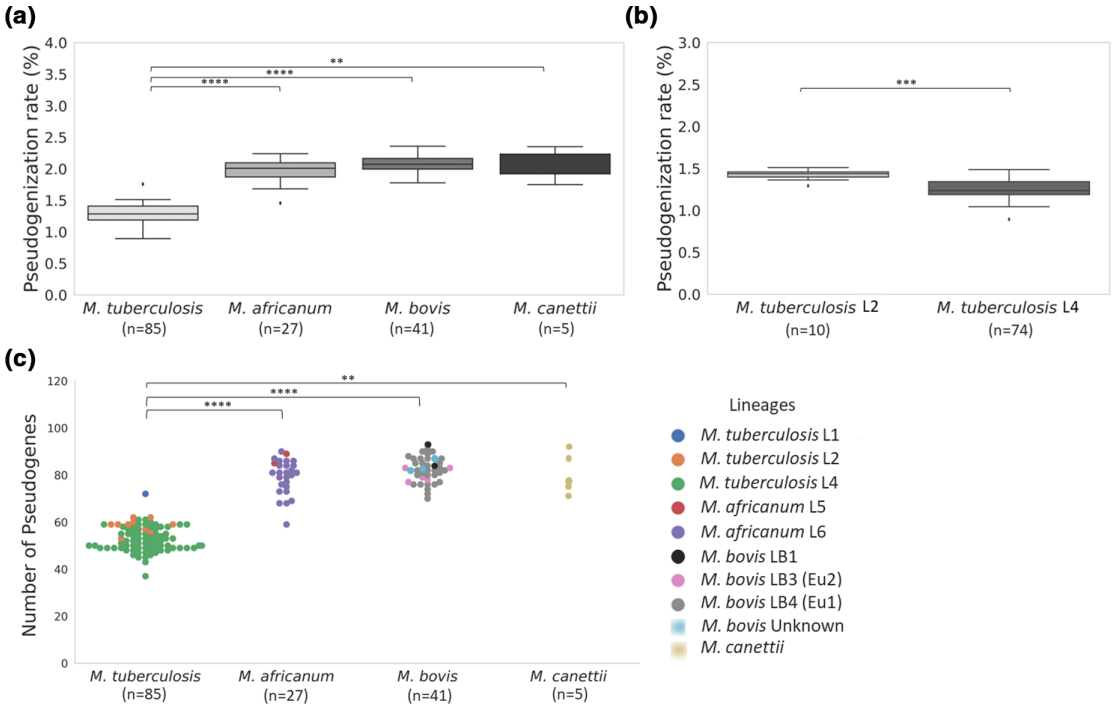

Fig. 1. Comparison of pseudogenization rates among *Mycobacterium tuberculosis* complex (MTBC) species and *Mycobacterium canettii*. (a) Distribution of the pseudogenization rates among MTBC species and *M. canettii*. (b) Distribution of the pseudogenization rates among *M. tuberculosis* lineages 2 and 4. (c) Distribution of the number of pseudogenes/genome among MTBC lineages and *M. canettii*. Statistical analysis was performed using Kruskal-Wallis, followed by Dunn's test in GraphPad Prism version 6. ***P*≤0.01 ****P*≤0.001 *****P*≤0.0001.

of pseudogene formation in all MTBC species ($P \leq 0.001$), while internal stop codon was the least common event leading to pseudogenization (Fig. 3). In contrast to MTBC species, strains of *M. canettii* did not show a significant difference between the occurrence of frameshifts and incomplete gene events (Fig. 3d), which may indicate a stronger influence of recombination [13] or other events that lead to gene disruption in *M. canettii* genomes.

Pseudogenes as drivers of populational genetic variability in the MTBC and *M. canettii*

To evaluate if genomes have the same pseudogenes, we identified how conserved the pseudogenes were among the strains evaluated in this study. The 10262 redundant pseudogenes were identified as 879 unique pseudogenes (i.e. non-redundant) (Table S2). Of these, 19 pseudogenes are present in ≥90% of the strains, while the remaining 860 are variably present among the genomes (510 are present in between two and 141 genomes and 369 are singletons) (Fig. 4, Table S2 (available in the online Supplementary Material), Supplementary Methods). There is also a conservation pattern according to species and lineages (Fig. 4). Taken together, our findings suggest that pseudogenes are a source of genetic variability to the MTBC.

Contribution of pseudogenes to the pan-genome of MTBC

The pan-genome of the MTBC and *M. canettii* strains is composed of 5366 genes. Pseudogenes are not included in this pan-genome, but their full-length gene counterparts are. A total of 879 pseudogenes were variably present in MTBC and *M. canettii* (i.e. present in less than 90% of the studied strains), which means that 16.38% (879/5366) of the genes of the pan-genome are subjected to pseudogenization. Therefore, an important proportion of the gene pool of MTBC is prone to variable pseudogenization, i.e. while a gene locus contains a pseudogene in one strain, it may harbour its full-length gene version in another strain. This finding supports that the process of pseudogenization contributes to the gene diversity of the MTBC at the population level.

Protein functions affected by pseudogenization

Using COG categorization, the top five functional classes in which MTBC and *M. canettii* pseudogenes were classified into were: (S) unknown function (24.90%), (I) lipid transport and metabolism (12.46%), (K) transcription (11.20%), (L) replication, recombination and repair (6.95%), and (G) carbohydrate transport and metabolism (6.87%). The 'lipid transport and metabolism' (I) category is composed mainly of acyl-CoA dehydrogenases, carboxylesterases, (R)-hydratases, enoyl-coA hydratases, epoxide hydrolases, and hypothetical proteins (682/914; 76.61%). The 'transcription' (K) category includes TetR family transcriptional

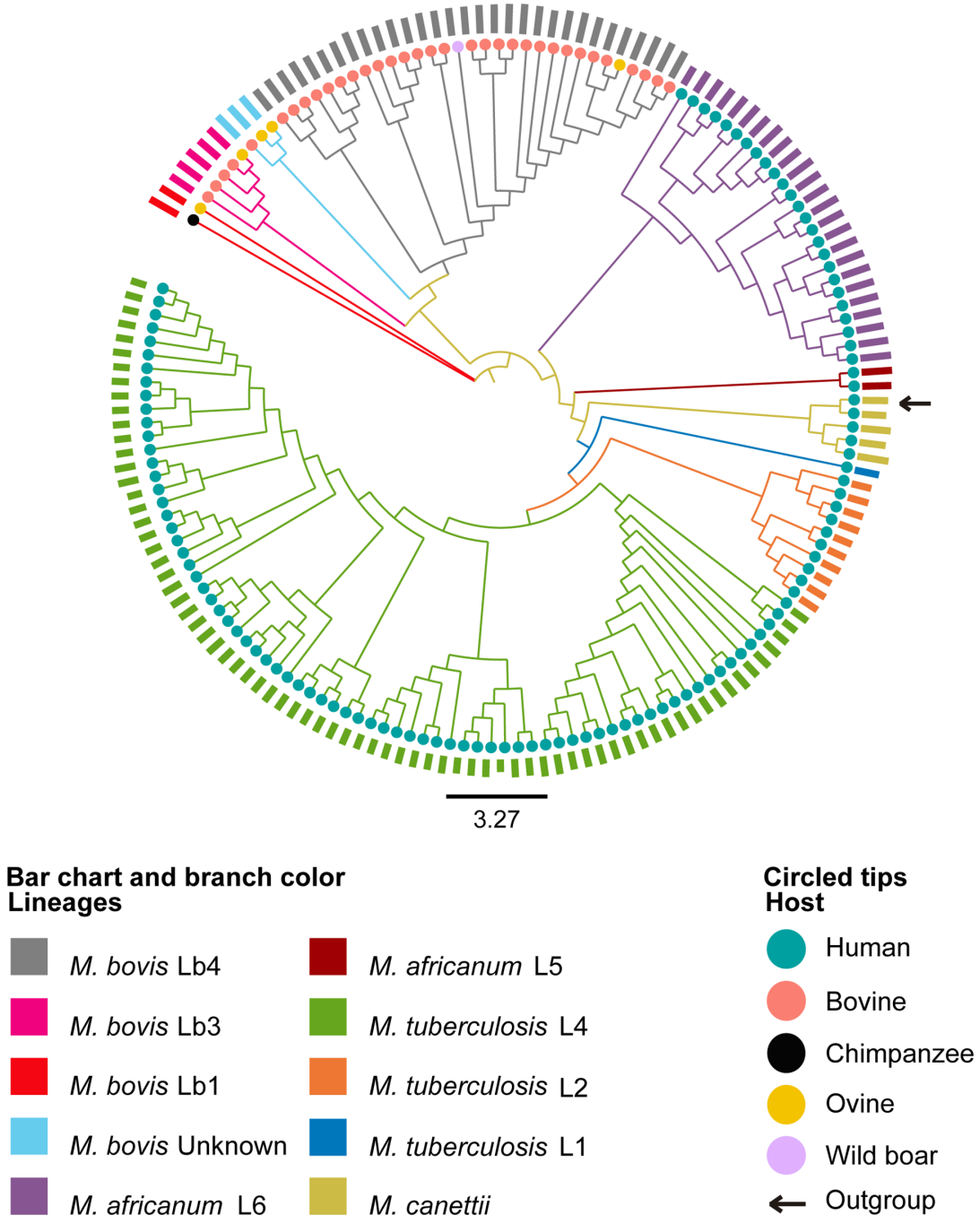

Fig. 2. Maximum likelihood phylogenetic tree based on core SNPs (single nucleotide polymorphisms) of 158 genomes of the *Mycobacterium tuberculosis* complex (MTBC) and *Mycobacterium canettii*. Bar charts indicate the number of pseudogenes of each genome. Coloured branches correspond to bacterial lineages and species. Circled tips correspond to host species from which each bacterial isolate was obtained according to NCBI (National Center for Biotechnology Information) metadata. *Mycobacterium canettii* CIPT 140010059 was used as outgroup. Phylogenetic tree was generated using IQ-Tree [41] with 1000 bootstrap replicates from a kSNP3 [40] matrix and graphically edited using Iroki [43]. Bootstrap replicas of main nodes are all ≥90%. Bar shows substitutions per nucleotide. The branch lengths of the tree were transformed using the cladogram option in FigTree [116] to improve visualization.

regulators, hydrogenase accessory proteins HypB, WXG100 family type VII secretion targets, cell filamentation proteins Fic, serine/threonine protein kinases, and hypothetical proteins (577/822; 70.19%). The 'replication, recombination and repair' (L) category is composed mainly of WXG100 family type VII secretion targets, serine/threonine protein kinases, exodeoxyribonucleases, LuxR family transcriptional regulators, and hypothetical proteins (375/510;73.53%).

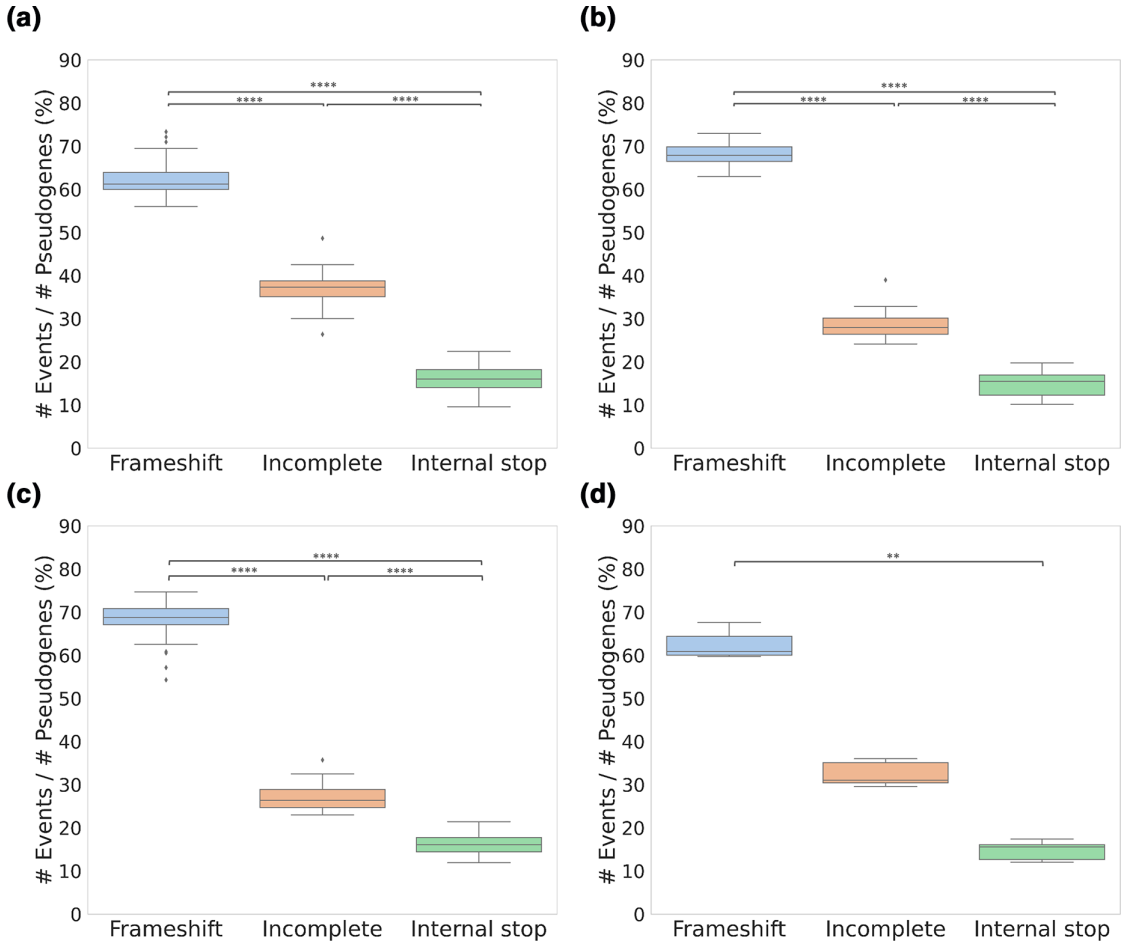

Fig. 3. Occurrence of events leading to pseudogenization (gene disruption) in the *Mycobacterium tuberculosis* complex (MTBC) and *Mycobacterium canettii*. The y-axis is the number of events divided by the number of *in silico* predicted pseudogenes in each genome. (a) *Mycobacterium tuberculosis*, (b) *Mycobacterium africanum*, (c) *Mycobacterium bovis*, (d) *M. canettii*. Statistical analysis was performed using Kruskal-Wallis, followed by Dunn's test in GraphPad Prism version 6. $^{**}P \leq 0.01$ $^{***}P \leq 0.001$ $^{****}P \leq 0.0001$.

COG enrichment analysis was performed by comparing the number of pseudogenes in each COG category against the entire COG-categorized proteome (genes) of the same species (Fig. 5a). The COG category commonly enriched among MTBC ecotypes was (D) cell cycle control, cell division, chromosome partitioning, composed mostly of the cell filamentation protein Fic and the EccC (164/199; 82.41%). *Mycobacterium tuberculosis* and *M. bovis* had only one additional commonly enriched category, the category (T) signal transduction mechanisms, composed mostly of WXG100 family type VII secretion targets, serine/threonine protein kinases, anti-anti-sigma factors, MMPL family RND transporters, adenylyl cyclases, and hypothetical proteins (301/353; 85.27%). In contrast, *M. tuberculosis* and *M. africanum* had four other enriched categories in common, the categories (F) nucleotide transport and metabolism, (N) cell motility, (O) post-translational modification, protein turnover and chaperones, and (P) inorganic ion transport and metabolism. The category F is mainly composed of anthranilate phosphoribosyltransferases, glycerol kinases, and hypothetical proteins (240/308; 77.92%). The category N is composed mostly of mammalian cell entry and hypothetical proteins (272/284; 95.77%). The category O includes hydrogenase accessory protein HypB, chaperone HtpG, membrane-anchored mycosin, peptidases, and hypothetical proteins (248/280; 88.57%). And the category P is composed mostly of adhesion proteins, MFS (major facilitator superfamily) transporters, and hypothetical proteins (275/37; 72.94%).

Mycobacterium bovis and *M. africanum* had two additional COG categories in common, the category (G) carbohydrate transport and metabolism, composed mostly of proteins of MFS transporters, proteins involved in glycosylation, NADP-dependent phosphogluconate dehydrogenases, and hypothetical proteins (358/504; 71.03%); and the category (H) coenzyme transport and metabolism, composed mostly of 8-amino-7-oxononanoate synthases, cobalt-precorrin-6A reductases, oxidoreductases, sulfotransferases, and hypothetical proteins (289/367; 78.75%). Interestingly, the categories (K) transcription and (L) replication, recombination and repair were enriched in *M. tuberculosis* alone. Collectively, these results indicate that certain functional classes

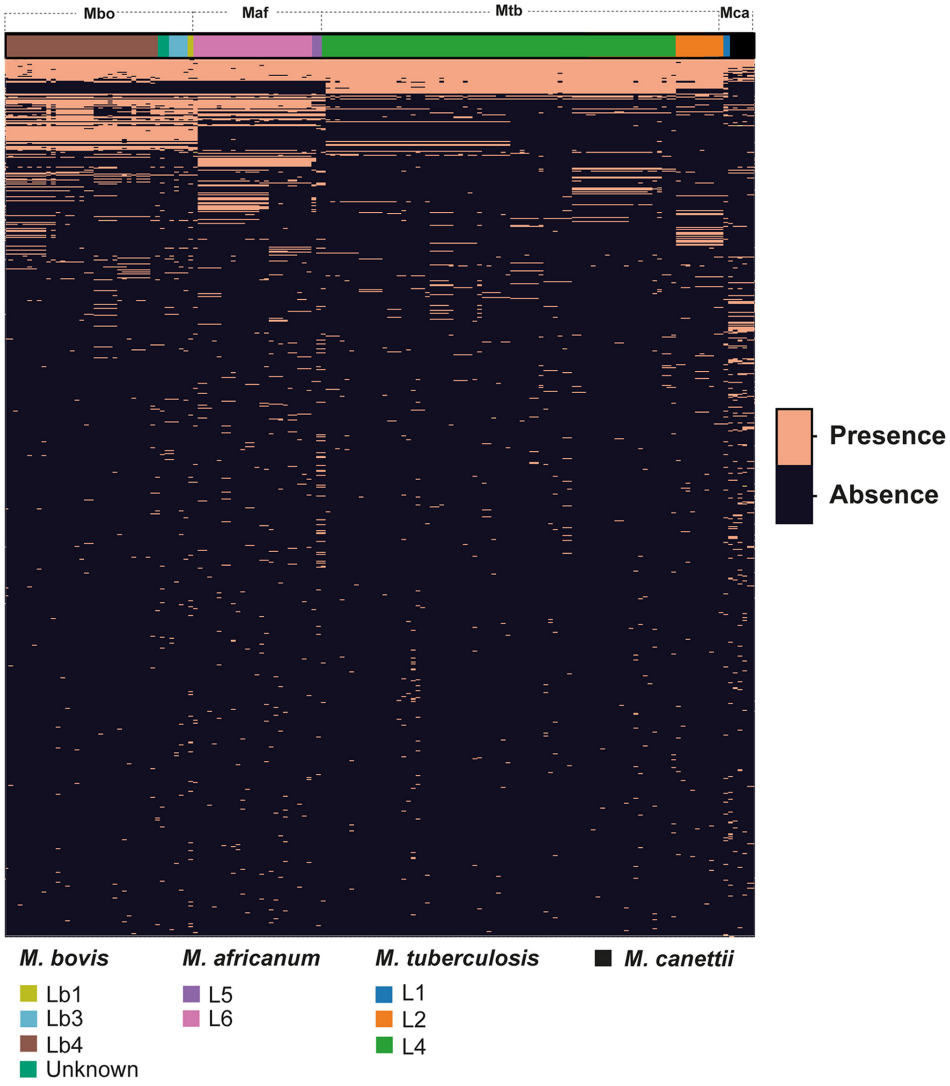

Fig. 4. Conservation of *in silico* predicted pseudogenes among the *Mycobacterium tuberculosis* complex (MTBC) and *Mycobacterium canettii* strains. The heatmap shows the distribution of the 10 262 *in silico* predicted pseudogenes among 879 loci of all strains. Copper-orange ('1'): pseudogene is present in the genome, and dark purple ('0'): pseudogene is absent. Pseudogene presence or absence was organized in a matrix with '1' and '0' and transformed into a heatmap using the Seaborn library [49], ordered from the most conserved to the least conserved pseudogenes. Mbo: *Mycobacterium bovis*, Maf: *Mycobacterium africanum*, Mtb: *M. tuberculosis*, Mca: *M. canettii*.

of MTBC and *M. canettii* are more likely to contain pseudogenes, and a few are species-specific; these are considered hotspots of gene remodelling.

COG categories were further evaluated according to bacterial lineages (Fig. 5b). *Mycobacterium tuberculosis* L1 and L2 have more pseudogenes mainly in the category I (lipid transport and metabolism) compared to *M. tuberculosis* L4 (Fig. 5b), which is likely responsible for the higher number of pseudogenes observed in the L1 and L2 lineages. Finally, there is an overall increase in the number of pseudogenes in all COG categories of *M. bovis* and *M. africanum* compared to *M. tuberculosis* L4, except for categories L (replication, recombination, and repair) and V (defence mechanisms), likely influencing their higher pseudogenization rate (Fig. 5b).

Using STRING database, it is also possible to observe functional protein associated networks among the detected pseudogenes with known functional annotation (i.e. gene names), regardless of whether they are pseudogenized simultaneously or not in the same strain (Fig. S5). These include but are not restricted to: genes of PDIM metabolism, serine/threonine-protein kinases (*pkn*), ESX-1 genes, cation-transporting P-type ATPases (*ctp*), *mce* genes, *glpk* gene, toxin-antitoxin systems (*vapC*), genes of sulfolipid metabolism, transcriptional regulators, among others (Fig. S5). A complete list of corresponding Rv numbers of detected non-redundant pseudogenes is provided in Table S2.

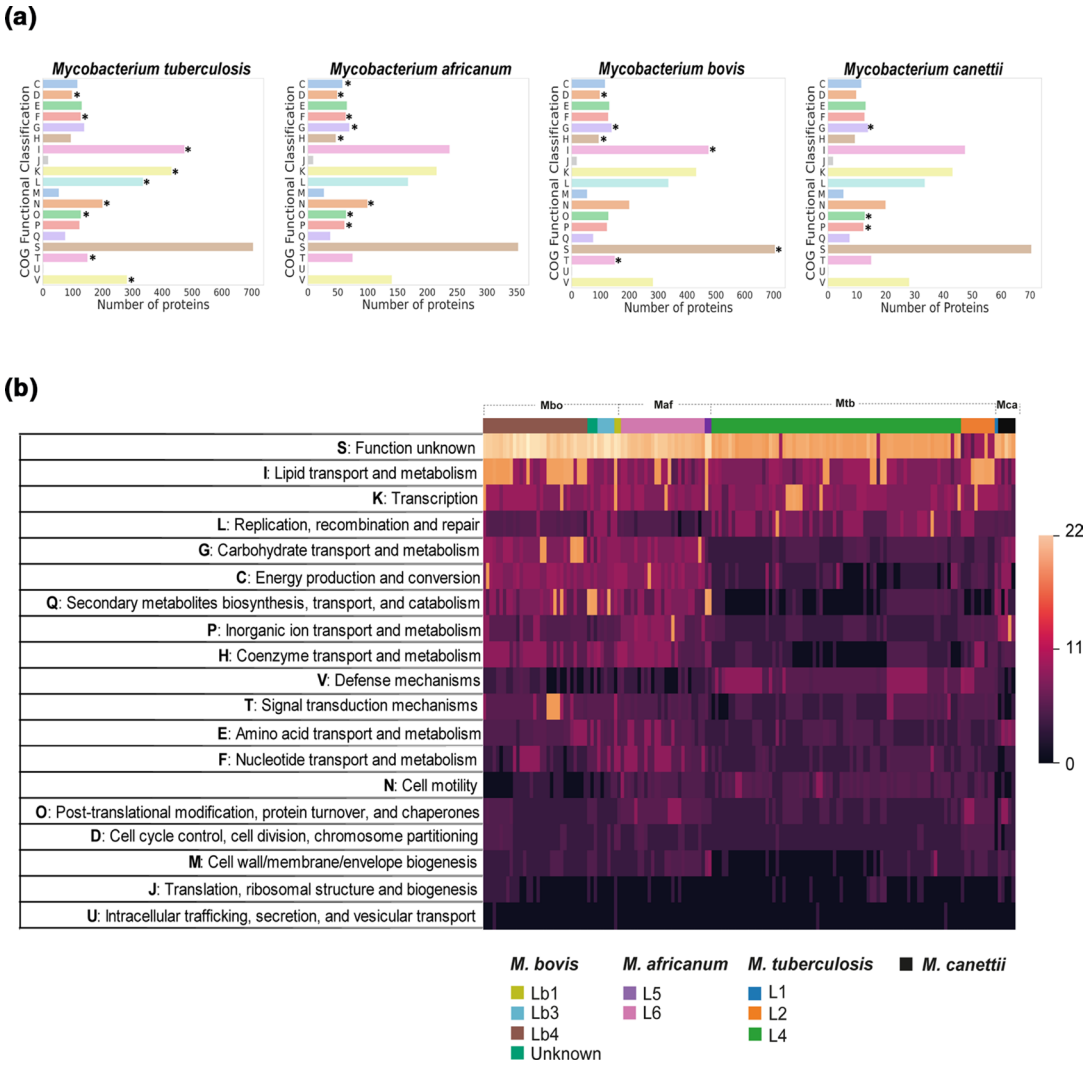

Fig. 5. Cluster of Orthologous Groups (COG) of *in silico* predicted pseudogenes of the *Mycobacterium tuberculosis* complex (MTBC) and *Mycobacterium canettii*. (a) COG enrichment analysis. A total of 70.34% of the pseudogenes and from 59.14–75.22% of the four species whole proteome dataset were successfully annotated using COG in EggNOG. The whole genome and full-length protein counterpart classification are available in Table S3. *Categories that appeared significantly enriched compared to COGs of the entire proteome of the same species, in a COG enrichment analysis ($P \leq 0.05$ and one degree of freedom using Chi-squared Test). (b) *In silico* predicted pseudogenes by COG functional classification among strains. The heatmap shows the distribution of 7337 pseudogenes with BER hit and COG classification. The remaining 3094 pseudogenes had no BER hit or COG classification. It is a non-exclusive analysis, i.e. a pseudogene can be assigned into one or more COG categories. The colour bar score ranges from absence ('0') to a maximum number of pseudogenes (22) found in a COG category in one strain. Mbo: *Mycobacterium bovis*, Maf: *Mycobacterium africanum*, Mtb: *M. tuberculosis*, Mca: *M. canettii*.

Antimicrobial resistance and tolerance

Frameshift of selected genes is a mechanism associated with phenotypes of antimicrobial resistance or tolerance. As expected, we found eight genes previously associated with antibiotic resistance carrying frameshifts in selected *M. tuberculosis* genomes: Rv3083, *ethA* (Rv3854c), *gid* (Rv3919c), Rv2752c, *mmpL5* (Rv0676c), Rv0678, *pncA* (Rv2043c), and *tlyA* (Rv1694). Three resistance-related genes were also detected with a frameshift in few *M. bovis* strains: *rpoB* (Rv0667), *eis* (Rv2416), and *rpoA* (Rv3457c); two *in M. africanum*: Rv2752c and Rv3083; and none in *M. canettii*.

We also uncovered a novel frameshift mutation potentially associated with resistance or tolerance to linezolid. Accordingly, six *M. tuberculosis* strains, including the H37Rv, carry a frameshift mutation at the 2/3 portion of the *rlmN* gene (Rv2879c). This gene is a putative 23S rRNA (adenine (2503) – C (2))-methyltransferase responsible for methylating the C2 position of the adenosine A2503 of the 23S rRNA, which is located at the peptidyl transferase center of the ribosome. Mutations in *rlmN* have been associated with resistance to a range of ribosome-targeting antibiotics in other bacteria [69]. Because annotated pseudogenes carry no

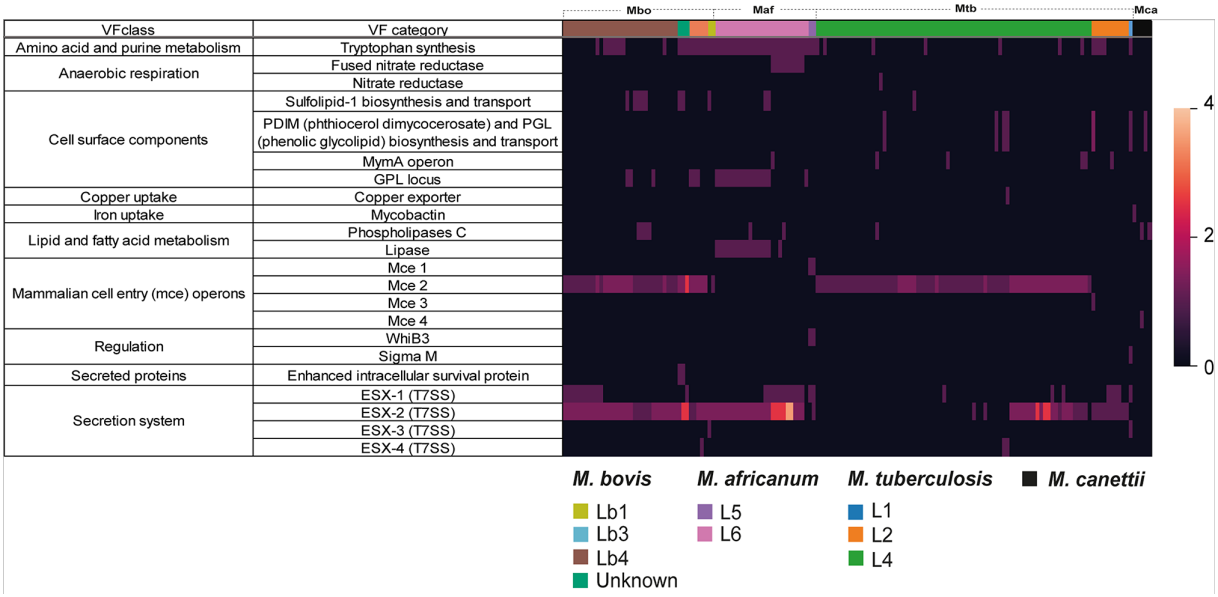

Fig. 6. Distribution of *in silico* predicted pseudogenes in virulence factor (VF) categories among the *Mycobacterium tuberculosis* complex (MTBC) and *Mycobacterium canettii* strains. The heatmap shows the distribution of the 535/10262 (5.21%) pseudogenes among the bacterial strains identified as VFs. They were identified among 50 virulence factor-related genes and concatenated into 22 categories. *Mycobacterium tuberculosis* (Mtb) showed the highest number of pseudogenized VF-related genes (*n*=27), followed by *M. bovis* (Mbo, *n*=19), *M. africanum* (Maf, *n*=16) and *M. canettii* (Mca, *n*=4), suggesting either a sample-size effect or a species association (Table S4). Pseudogene presence or absence was organized in a matrix with '1' and '0', then presence was counted by category/genome, creating a heatmap using the Seaborn library [49]. The colour bar score ranges from absence ('0') to a maximum number of pseudogenes (four) found in a category in one strain ('1'). Data according to each gene can be found in Fig. S6 and Table S4.

gene ID, it is possible that genomic studies evaluating linezolid resistance in *M. tuberculosis* may miss potential mutations in this gene depending on the reference H37Rv strain being used to map and call variants. It is also possible that mutations in *rlm*N are associated with antibiotic tolerance and not necessarily resistance, which warrants further studies.

Antimicrobial tolerance is a major challenge in the treatment of tuberculosis. In this study, frameshifts in the *glpk* gene (Rv3696c; glycerol kinase), previously associated with tolerance [70], have been detected in selected genomes from all three MTBC species. Pseudogenization of genes associated with sulfolipid biosynthesis and transport (*pks2*, *papA1, mmpL8*), PDIM biosynthesis (*ppsC*, *ppsD*), fatty acid biosynthesis and oxidation (*fadE* and *fadD* genes, *acrA1*), and several other genes involved in lipid metabolism were also detected (Table S2). These findings suggest modulation of cell wall composition, which in turn may lead to increased tolerance to environmental stressors, including antibiotics.

Pseudogenization of virulence factors (VF)

Fifty VF genes (grouped into 22 functional categories and representing 535 of the 10 262 redundant pseudogenes) were detected among pseudogenes of the analysed strains (Fig. 6, Fig. S6, Table S4). The most conserved VF-related pseudogene is *mce2B* (Rv0590) (mammalian cell entry operon–related protein), present in 74 strains, followed by *esxC* (Rv3890c), present in 65 strains (Table S4). The top six VF categories prone to pseudogenization were: ESX-2/T7SS (186 of 535 redundant pseudogenes; 34.76%), Mce2 (160/535; 29.91%), tryptophan synthesis (57/535; 10.65%), ESX-1/T7SS (33/535; 6.17%), GPL (glycopeptidolipids) locus (22/535; 4.11%), and lipase (16/535; 2.99%). Together, these categories represent 88.56%(474/535) of the VF genes prone to pseudogenization (Fig. 6).

Lineage or group-specific patterns among VF categories were observed (Fig. 6, Fig. S6, Table S4). Operons of the type VII secretion were pseudogenized mainly in *esx*, *esp, mycP* and *ecc* genes of ESX-1, ESX-2, ESX-3, and ESX-4 (Fig. 6, Fig. S6, Table S4). Considering these systems, the ESX-2 and ESX-1 are the most affected among MTBC strains, while intact in *M. canettii* (Fig. 6, Fig. S6, Table S4). The Mce2 operon genes *mce2B*, *mce2D* (Rv0592) and *mce2E* (Rv0593) were pseudogenized in 74, 47 and 39 strains, respectively. The *mce2B* appeared pseudogenized in all *M. tuberculosis* L4 strains, *mce2D* is pseudogenized in many *M. bovis* (*n*=19/41) and *M. tuberculosis* L4 strains (*n*=28/74), and *mce2E* is pseudogenized in 39 of the 41 *M. bovis* strains (Fig. S6 and Table S4). Genes of Mce operons 1 (*mce1C*; Rv0171), 3 (*mce3F*; Rv1971), and 4 (*mce4F*; Rv3494c) were pseudogenized in only up to two strains (Fig. S6 and Table S4). Noteworthy, the Mce3 operon is deleted in *M. bovis* and *M. africanum* strains [71, 72].

The *trpD* gene (Rv2192c; the sole representative of the tryptophan synthesis category) is pseudogenized in representatives of all MTBC ecotypes and is intact in *M. canettii* genomes. The putative glycosyltransferases 3 (*gtf3*) Rv1524 and Rv1526c of the GPL locus are pseudogenized in 16 *M. africanum* L6 strains and in five *M. bovis* strains, respectively. The *pks2* (Rv3825c), a gene of the sulfolipid biosynthesis and transport, was also pseudogenized in five *M. bovis* strains and in two *M. africanum* L6, while *papA1* (Rv3824c) was pseudogenized in two *M. bovis* strains and in one *M. tuberculosis* L4 (Fig. 6, Fig. S6, Table S4).

Interestingly, the gene *lipF* (Rv3487c), encoding a dual carboxylesterase and phospholipase, is pseudogenized in 16 out of the 25 strains of *M. africanum* L6, while the *narX* gene (Rv1736c), a fused nitrate reductase, is pseudogenized in nine strains of *M. africanum* L6 (Table S4). The presence of these pseudogenes was almost exclusionary of one another; only one L6 strain has both *lipF* and *narX* as pseudogenes, and one L6 strain does not carry *lipF* nor *narX* pseudogenes. In addition, strains of *M. africanum* L6 carrying the *gtf3* (Rv1524) pseudogene have intact *narX*, while those carrying a *narX* pseudogene have an intact *gtf3*. Thus, 15 *M. africanum* L6 strains carry both *gtf3* and *lipF* pseudogenes. The gene *narX* was not found as a pseudogene in any other MTBC or *M. canettii* strain (Fig. S6 and Table S4).

Finally, the phospholipase locus *plcABCD* of *M. bovis* has been deleted by the recombination of an IS6110 element [73]. Consequently, it was believed that *M. bovis* would carry only one phospholipase, the *plcD* gene (Rv1755c). Our study shows that *plcD* is a pseudogene in four *M. bovis* strains and two *M. canettii* strains (Table S4). Whib3, a vital redox sensor that allows survival amid reactive oxygen and nitrogen species [74], was found pseudogenized in both strains of *M. africanum* L5.

Gene essentiality

The essentiality of the pseudogenes for growth *in vitro* was evaluated using previous work from DeJesus *et al.* [55] and Gibson *et al.* [56]. As expected, only few pseudogenes (30 of the 879 non-redundant/unique pseudogenes; 3.41%) were considered essential for *in vitro* growth of *M. bovis* and/or *M. tuberculosis* (Table S5). Interestingly, more pseudogenes that can lead to *in vitro* growth advantage were detected (79/879; 8.99%) than pseudogenes essential for growth (Table S5).

Annotated pseudogenes are expressed as mRNA in *M. tuberculosis* H37Rv

We investigated whether the pseudogenes are expressed as mRNA in the model organism *M. tuberculosis* H37Rv. The genome of *M. tuberculosis* H37Rv contains 92 pseudogenes (NC_018143.2; Table S6). A total of 64.13% (59/92) of these pseudogenes were detected in the transcriptome, with a read depth varying from 10 to 1453 reads among the six replicates. Other 13 (14.13%) pseudogenes showed a mean read depth between one and 10 reads among the six replicates. Finally, 20 pseudogenes presented a read depth of zero (Table S7), implying the absence of transcription in the study conditions. Thus, as expected, the transcription machinery can fully transcribe most pseudogenes.

Proteomic analyses reveal translation of pseudogenes in *M. tuberculosis* H37Rv

To evaluate if these pseudogenes are translated into proteins, a published proteome of *M. tuberculosis* H37Rv was reanalysed [62]. Surprisingly, 13/92 (14.13%) pseudogene loci were found translated in the conditions tested (Table S8). These proteins had a peptide coverage ranging from 2.3–59.3% (average 17.6%), and the probability of each detected fragment being truly from the respective protein was >97%. Analysis of the positions of the detected peptides along the length of the protein revealed two (2/13) proteins with peptides detected downstream/C-terminal of the pseudogenization lesion (i.e. the frameshift mutation) (Fig. S7), suggesting that these proteins are translated as predicted by the PGAP gene finder and may represent protein sequence variation over the evolutionary course of MTBC.

We then corrected the frameshift or internal stop codon of 11 out of the 13 proteins detected by MS/MS to verify peptides downstream/C-terminal of the pseudogenization lesion. No peptides corresponding to the 'corrected' portion of the proteins were detected. Therefore, apart from the two proteins described above, it was not possible to determine if the remaining translated proteins are expressed as disabled or corrected forms.

DISCUSSION

Our study is the first to comprehensively analyze *in silico* predicted pseudogenes of the MTBC and *M. canettii*. Results show significant variability concerning the pseudogenization rate among analysed ecotypes and lineages, with most of the pseudogenes (~98%) variably present among strains, highlighting functional hotspots of mutational lesions associated with pseudogenization. Moreover, we show that these alterations in gene sequences contribute to the genetic diversity of MTBC at the population level, causing important variations in virulence factors and other genes of the metabolism and antimicrobial resistance/tolerance.

Why do the rates of *in silico* predicted pseudogenes vary among species of the MTBC?

MTBC's evolution is driven by purifying and background selections, selective sweeps, and transmission bottlenecks [75–77]. Population bottlenecks occur during host-to-host transmission and influence the emergence of new mutations through genetic

drift [75, 78–82], which causes population reduction and relaxation of natural selection, limiting population diversity [75, 78]. These mechanisms have been suggested to increase permissiveness to deleterious mutations, resulting in pseudogenization [32, 33]. Thus, accumulative population bottlenecks and the founding of new bacterial populations (i.e. founder effect) [77, 82, 83] can explain why *M. africanum* and *M. bovis* have higher pseudogenization rates when compared to *M. tuberculosis* L4, as also suggested previously [84]. Similarly, the high transmissibility rate and antibiotic resistance of *M. tuberculosis* L2 [85–87] in populous countries, mainly driven by Beijing sublineages [88–90], likely led to increased genetic drifts and, consequently, higher rates of pseudogenization when compared to *M. tuberculosis* L4. Finally, the high pseudogenization rates found in *M. canettii* can be explained by their more recombinogenic genome [13, 30], as these mechanisms can result in disruptive changes in the genetic structure and organization.

The higher pseudogenization rate of *M. bovis* compared to *M. tuberculosis* corroborates previous work showing its genomic decay compared to the human-adapted pathogen, defined by the deletion of genomic regions [91]. Because *M. bovis* infects a broader range of host species, this genomic decay contrasts with other bacteria in which gene loss is frequently associated with host specialization [92, 93]. However, since it is not clear if the genetic mechanisms of frameshift and nonsense mutations always lead to fixed gene inactivation (i.e. the classical definition of a pseudogene) in MTBC, the *in silico* detection of these truncated sequences may not necessarily imply permanent genomic decay. Instead, gene content modulation through these genetic mechanisms may allow higher metabolic versatility to survive in different environments and, perhaps, host species.

Genetic events leading to pseudogenization: the importance of indels

As observed in other bacteria [34], frameshifts are the leading cause of gene disruption in MTBC and *M. canettii*. Thus, our results underscore indels and consequent frameshifts as drivers of evolution and genetic variability in the MTBC. IS elements also play an important role in pseudogene emergence; their introduction into genomes may lead to the disruption of genes, generating *in silico* predicted pseudogenes classified as 'incomplete genes'. The latter has greater chance of fixation than frameshifts. Certain frameshifts have been shown to reverse in MTBC [70, 94, 95]. Lastly, the low number of *in silico* predicted pseudogenes with internal stop codon is likely a reflection of the high GC content of mycobacteria, hampering the AT-rich stop codon emergence [34]. Therefore, in the absence of HGT, indels and IS are crucial promoters of genetic variability in the MTBC.

Variable distribution of pseudogenes among strains contributes to genetic diversity at the population level

Only 19 pseudogenes were found in ≥90% of the strains; the vast majority of the pseudogenes (n=860/879) are not conserved among the analysed genomes. Thus, while a gene is pseudogenized in one strain, the other has a full-length gene counterpart in its equivalent position. There is also a conservation pattern according to lineages and ecotypes, underscoring the potential role of pseudogenization in determining phenotype. Given this variable pseudogene distribution, gene loci under pseudogenization corresponded to ~16% of the MTBC pan-genome. Collectively, these results indicate that pseudogenes are an important source of genetic diversity for the MTBC.

The reasons for the variable distribution of 860 pseudogenes are likely a combination of variation in the number and presence of IS elements among strains [96], the possibility of gene phase variation [70, 94], or not enough evolutionary time for pseudogene fixation. Phase variation has been reported with the glycerol kinase gene (*glpk*) [70, 94], and thus, it is possible that some of the detected pseudogenes are products of such genetic event, and in theory, may reverse to their original state. This mechanism of reversible gene silencing has been recently highlighted in *M. tuberculosis* [95]. In addition, reversion of nonsense mutations has also been described in bacteria [97]. This possibility of pseudogenes' reversal may unprecedently change the way we perceive genetic diversity in MTBC at the individual bacterium and population levels, and further studies are warranted.

Protein functions affected by pseudogenization in tuberculous mycobacteria

Despite variation in pseudogenes conservancy, there is an overall similarity in the type of functional categories carrying genes under pseudogenization among the studied ecotypes. The top five categories account for half of the pseudogenes and can be considered hotspots of pseudogenization. The 'unknown function' is the category with the highest number of detected pseudogenes in MTBC and *M. canettii*, except for part of *M. tuberculosis* L2 in which the 'lipid transport and metabolism' category stands out. An interesting finding was the enrichment of the 'transcription' and 'replication, recombination and repair' categories in *M. tuberculosis*. Pseudogenes within these categories include various types of transcriptional regulators, serine/threonine protein kinases, methyltransferases, and Fic-domain proteins, suggesting variable loss of functions with implications in gene expression and post-translational modifications. Loss of function of similar regulators due to indels and non-sense mutations have been described in other bacteria as means to adapt to changes in the microenvironment and quorum-sensing [98]. The category 'carbohydrate transport and metabolism' enriched in *M. bovis*, *M. africanum* and *M. canettii* also contains pseudogenes of proteins that participate in post-translational glycolisation. Glycolysation of proteins and lipids of MTBC exerts essential functions from cell wall composition to modulation of host immune response [99]. Their pseudogenization is possibly linked to the development of distinct *M. bovis* and *M. africanum* phenotypes.

The pseudogenization of virulence factors

Our study underscores the importance of loss of function mutations to MTBC virulence plasticity. The ESX-2 system was the most affected VF category; however, its function is poorly described as this operon is not required by *M. tuberculosis* for *in vitro* growth or virulence in mice [100]. While the role of the ESX-1 system is well known [100], the function of the pseudogenes detected herein (*espJ* and *espK*) is less understood, but EspK has been implicated in the processing of EspB, CFP-10, and ESAT-6, important mycobacterial antigens [101].

Few genes of the Mce2 operon and the sulfolpid metabolism (*pks2* and *papA1*) are pseudogenized in certain *M. tuberculosis*, *M. africanum* and/or *M. bovis* strains. Mce2 knock-out leads to increased accumulation of sulfolipids in the cell wall of *M. tuberculosis* [102], but since the structure of the Mce2 transporter is unknown, the impact of the pseudogenization of *mce2B* and *mce2D* on the transporter's function cannot be predicted. Interestingly, *M. tuberculosis* mutants of *pks2* are unable to produce sulfolipids [103]. Mutations in the PhoPR two-component system in strains of *M. bovis, M. africanum,* and *M. tuberculosis* H37Ra have also been associated with a lack of sulfolipid-1 production [104–106]. Moreover, the gene *lipF*, part of the PhoPR regulon, was pseudogenized in certain *M. africanum* L6 genomes. This gene is upregulated under acidic pH, its protein is localized in the cell wall, and possesses carboxylesterase and phospholipase C activities [107]. Collectively, these results suggest that PhoPR mutations in *M. bovis* and *M. africanum* [108] allowed pervasive erosion of regulon-associated genes (*pks2* and *lipF*) and also those related to sulfolipid metabolism (i.e. Mce2 genes and *pks2*).

Pseudogenes or actual genes?

This study found two pseudogenes predicted as translated by the PGAP in *M. tuberculosis* H37Rv. One of these proteins is a HAD hydrolase. HAD hydrolases are paralogous genes whose frameshifts likely contribute to sequence variation and not inactivation, implying this protein is functional. The second detected pseudogene belongs to the GntR family transcriptional regulator with a frameshift at the C-terminus. The predicted protein varies in amino acid sequence content after the mutational lesion, but in size by only seven amino acids. As some proteins can maintain their functional activity with a loss of <10% of their C-terminus [34], it is also possible that this protein is functional despite the frameshift and some degree of amino acid sequence variation.

Among the remaining eleven expressed proteins, one stands out as possibly functional: the anthranilate phosphoribosyltransferase (TrpD), a virulence factor part of the tryptophan biosynthesis pathway. Its frameshift causes a protein size variation of only six amino acids at the C-terminus. Strains of *M. tuberculosis* knocked out of *trpD* fail to cause disease in mice [109] and lead to tryptophan auxotrophy. Thus, it is likely that this frameshifted protein is also functional. In addition, even though PE/PPE pseudogenes were not included in the analyses, it is possible that many of their mutations are not sequencing or assembly errors. Considering the sequence and size variation among PPE/PE paralogous genes [110], their frameshifts or internal stop codons may be a source of genetic variability generating functional genes. Therefore, these results suggest that at least part of what is predicted as pseudogenes *in silico* is a source of metabolic plasticity rather than gene inactivation. While the PGAP pipeline can predict true pseudogenes (i.e. completely inactive genes) later experimentally proven, few predicted pseudogenes may still generate functional proteins. Further studies should be conducted to evaluate the function of these truncated proteins to elucidate the possibility of selective sweeps and fixation of novel gene variants associated with metabolic plasticity of the MTBC.

Limitations of this study

We cannot neglect the possibility of errors arising from sequencing, assembly, and/or gene prediction that may have led to the misidentification of pseudogenes by PGAP. However, Illumina sequencing has a very low error rate for indels [111], varying from $2.8{\cdot}10^{-6}$ to $4.9{\cdot}10^{-6}$ errors per base only [112]. Genomes from all other sequencing platforms, which have higher sequencing error rates in homopolymeric tracts, were not included in this study. Thus, sequencing errors are likely minor to the conclusions drawn from this study. It is also important to highlight that assemblies work with consensus sequences; thus, we cannot account for heterogeneous indels within the same locus.

The use of short-read sequences to assemble genomic areas containing DNA repeats is a long-standing computational challenge resulting in assembly ambiguities [113]. Thus, we excluded sequences that are known to be repetitive (PE/PPE genes, mobile genetic elements – insertion sequences, transposases, phage, integrases, and maturases) from the initial dataset of pseudogenes. Although it is possible that the detected lesions in these areas are true mutations, confirmation with other technologies will be required to rightly assert this in the future. Noteworthy, with or without these genes, observed differences in the pseudogenization rates between ecotypes were unchanged (data not shown). Nevertheless, shorter repeats may still be present in some of the analysed genes, and results, particularly of singletons, should be interpreted with caution until we know the true impact of repeats' contraction and expansion in genes of the MTBC.

We also excluded 861 non-redundant pseudogenes from our dataset because we could not identify their full-length gene counterparts (Supplementary Methods). Due to the high conservancy of MTBC genomes, we do not expect full-length genes to have been eliminated throughout evolution. The detection of full-length gene counterparts is a difficult task, especially for pseudogenes

that are incomplete and very fragmented, and this finding can be a limitation of the chosen methodology. However, it is also possible that these are mistaken pseudogene predictions on alternative strands, short pseudogene predictions upstream of a gene with an erroneous start site annotation, pseudogenes predicted in intergenic regions, or misassemble of repetitive areas resulting in inaccurate pseudogene predictions. Improvement of sequencing technologies, assembly methodologies, and gene prediction algorithms are needed to overcome barriers in the study of pseudogenes.

Without experimental studies, it is not possible to accurately determine which sequences evaluated in this study fit the classical definition of a 'pseudogene'. This study and others unveiled the possibility of frameshift or nonsense mutation reversal, protein neofunctionalization, and mutations that have no impact on protein function in MTBC. By definition, these sequences could not be called pseudogenes. To avoid further confusion, we kept the name 'pseudogene' or '*in silico* predicted pseudogenes' for all sequences evaluated herein, but researchers should keep in mind that these are called pseudogenes according to what PGAP defines as a pseudogene using solely *in silico* information.

Final considerations

Our study reports that *in silico* predicted pseudogenes of the MTBC are a source of genetic variability at the population level. The loss of function and possible phase variation or protein neofunctionalization bring phenotypic plasticity to this clonal group in the absence of horizontal gene transfer. The latter two possibilities need to be further evaluated by experimental work, albeit reports in the literature already support some of these phenomena [70, 94, 95]. Thus, mutations detected by pseudogene algorithms may elucidate genetic mechanisms by which MTBC acquires new adaptive traits.

While the effects of purifying and background selections in MTBC have been described [75, 76], the consequences of genetic drift caused by population bottlenecks are less understood. We suggest that at least part of the MTBC's *in silico* predicted pseudogenes result from population bottlenecks and genetic drift that lead to a relaxation of natural selection [76]. If the predicted pseudogenes are a source of gene variants, these, in turn, may be subjected to selective sweeps as effective population size expands. Genes under phase variation due to loss of function mutations may also be under alternative evolutionary pressures. Therefore, different evolutionary forces may act on the identified *in silico* predicted pseudogenes that need further studies.

We developed and made a series of scripts available through GitHub to quantify and evaluate the conservancy of bacterial pseudogenes from genomes available in public databases. We also show that given the absence of gene IDs for pseudogenes, traditional transcriptome analyses will not report differentially expressed pseudogenes. Herein, we used pseudogenes predicted by PGAP, but other annotation platforms identify pseudogenes differently (or do not identify) and this fact should be taken into consideration when comparing the results of this study or when performing estimates of pan-genome. Similar pseudogene annotation limitations have been described in other bacterial species whose pseudogenes have been extensively studied [114, 115], highlighting the difficulties in accessing these genes and recognizing their importance. Notably, a few predicted pseudogenes were found to be translated into proteins; thus, we believe genome annotations of MTBC should be performed by combining proteomic and functional assays to correct for possible inconsistencies.

Funding information
NSC fellowships were funded by Sao Paulo Research Foundation (FAPESP, grant number 2017/20147-7) and CNPq (Brazilian Ministry of Science, grant number 140003/2019–3). CKZ and AHA fellowships were funded by FAPESP (grant number 2017/04617–3 and 2019/03232–6). TTSP fellowship was funded by CAPES (Brazilian Ministry of Education, grant number 88887.508739/2020–00). MFM fellowship was also funded by CAPES. Part of this work was funded by FAPESP (grant number 2016/26108–0). Partial graduate studies funding was provided by CAPES (Finance code 001). This work was funded by Morris Animal Foundation (grant number D17ZO-307).

Acknowledgement
We are thankful to the Institute for Genome Sciences of the University of Maryland for the Manatee software and CEFAP (Core Facility to Support Research) of the Institute of Biomedical Sciences, University of São Paulo, for core computer services.

Author contributions
N.S.C. designed experiments, developed computational algorithms, performed, and interpreted bioinformatic analyses, wrote, and revised the manuscript. T.T.S.P. and C.K.Z. provided bioinformatics support and revised the manuscript. A.H.A. developed computational algorithms. M.F.C. and A.Z. performed the proteomic analyses. J.S. performed phylogenetic analysis and ancestor state reconstruction. A.P.S. revised the manuscript and contributed to the final analyses and interpretation. A.M.S.G. designed experiments, supervised the work, wrote, and revised the manuscript. All authors have read, revised, and agreed to the submission of the manuscript.

Conflicts of interest
The authors declare no conflict of interest.

References

1. Brites D, Loiseau C, Menardo F, Borrell S, Boniotti MB, *et al*. A new phylogenetic framework for the animal-adapted *Mycobacterium tuberculosis* complex. *Front Microbiol* 2018;9:2820.
2. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, *et al*. A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region. *Nat Commun* 2020;11:1–11.

3. Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, *et al*. Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. *Microb Genom* 2021;7:1–14.
4. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin Immunol* 2014;26:431–444.
5. de Jong BC, Adetifa I, Walther B, Hill PC, Antonio M, *et al*. Differences between tuberculosis cases infected with *Mycobacterium africanum*, West African type 2, relative to Euro-American *Mycobacterium tuberculosis*: an update. *FEMS Immunol Med Microbiol* 2010;58:102–105.
6. de Jong BC, Antonio M, Gagneux S. *Mycobacterium africanum*--review of an important cause of human tuberculosis in West Africa. *PLoS Negl Trop Dis* 2010;4:e744.
7. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, *et al*. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in the Gambia. *J Infect Dis* 2008;198:1037–1043.
8. Gagneux S. Strain variation in the *Mycobacterium tuberculosis* complex: its role in biology, epidemiology and control. 1019. 2017. DOI: 10.1007/978-3-319-64371-7.
9. Prodinger WM, Indra A, Koksalan OK, Kilicaslan Z, Richter E. *Mycobacterium caprae* infection in humans. *Expert Rev Anti Infect Ther* 2014;12:1501–1513.
10. WHO. *Global tuberculosis report 2019*. WHO, 2019.
11. Olea-Popelka F, Muwonge A, Perera A, Dean AS, Mumford E, *et al*. *Zoonotic tuberculosis* in human beings caused by *Mycobacterium bovis*-a call for action. *Lancet Infect Dis* 2017;17:e21–e25.
12. Duffy SC, Srinivasan S, Schilling MA, Stuber T, Danchuk SN, *et al*. Reconsidering *Mycobacterium bovis* as a proxy for *zoonotic tuberculosis*: a molecular epidemiological surveillance study. *Lancet Microbe* 2020;1:e66–e73.
13. Chiner-Oms Á, Sánchez-Busó L, Corander J, Gagneux S, Harris SR, *et al*. Genomic determinants of speciation and spread of the *Mycobacterium tuberculosis* complex. *Sci Adv* 2019;5:eaaw3307.
14. Boritsch EC, Khanna V, Pawlik A, Honoré N, Navas VH, *et al*. Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria. *Proc Natl Acad Sci* 2016;113:9876–9881.
15. Bolotin E, Hershberg R. Gene loss dominates as a source of genetic variation within clonal pathogenic bacterial species. *Genome Biol Evol* 2015;7:2173–2187.
16. Brosch R, Gordon SV, Billault A, Garnier T, Eiglmeier K, *et al*. Use of a *Mycobacterium tuberculosis* H37Rv bacterial artificial chromosome library for genome mapping, sequencing, and comparative genomics. *Infect Immun* 1998;66:2221–2229.
17. Philipp WJ, Nair S, Guglielmi G, Lagranderie M, Gicquel B, *et al*. Physical mapping of *Mycobacterium bovis* BCG pasteur reveals differences from the genome map of *Mycobacterium tuberculosis* H37Rv and from M. bovis. *Microbiology* 1996;142 (Pt 11):3135–3145.
18. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, *et al*. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol* 1999;32:643–655.
19. 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.
20. Silva-Pereira TT, Ikuta CY, Zimpel CK, Camargo NCS, de Souza Filho AF, *et al*. Genome sequencing of *Mycobacterium pinnipedii* strains: genetic characterization and evidence of superinfection in a South American sea lion (Otaria flavescens). *BMC Genomics* 2019;20:1030.
21. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, *et al*. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci* 2002;99:3684–3689.
22. Gagneux S, Small PM. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis* 2007;7:328–337.
23. Galagan JE. Genomic insights into tuberculosis. *Nat Rev Genet* 2014;15:307–320.
24. Isaza JP, Duque C, Gomez V, Robledo J, Barrera LF, *et al*. Whole genome shotgun sequencing of one Colombian clinical isolate of *Mycobacterium tuberculosis* reveals DosR regulon gene deletions. *FEMS Microbiol Lett* 2012;330:113–120.
25. DeJesus MA, Sacchettini JC, Ioerger TR. Reannotation of translational start sites in the genome of *Mycobacterium tuberculosis*. *Tuberculosis* 2013;93:18–25.
26. Xiong X, Wang R, Deng D, Chen Y, Liu H, *et al*. Comparative genomics of a bovine *Mycobacterium tuberculosis* isolate and other strains reveals its potential mechanism of bovine adaptation. *Front Microbiol* 2017;8:2500.
27. Jia X, Yang L, Dong M, Chen S, Lv L, *et al*. The bioinformatics analysis of comparative genomics of *Mycobacterium tuberculosis* complex (MTBC) provides insight into dissimilarities between intraspecific groups differing in host association, virulence, and epitope diversity. *Front Cell Infect Microbiol* 2017;7:88.
28. Yang T, Zhong J, Zhang J, Li C, Yu X, *et al*. Pan-genomic study of *Mycobacterium tuberculosis* reflecting the primary/secondary genes, generality/individuality, and the interconversion through copy number variations. *Front Microbiol* 2018;9:1–12.
29. Dippenaar A, Parsons SDC, Sampson SL, van der Merwe RG, Drewe JA, *et al*. Whole genome sequence analysis of *Mycobacterium suricattae*. *Tuberculosis* 2015;95:682–688.
30. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, *et al*. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nat Genet* 2013;45:172–179.
31. Liu F, Hu Y, Wang Q, Li HM, Gao GF, *et al*. Comparative genomic analysis of *Mycobacterium tuberculosis* clinical isolates. *BMC Genomics* 2014;15:469.
32. Kuo CH, Ochman H. The extinction dynamics of bacterial *pseudogenes*. *PLoS Genet* 2010;6:e1001050.
33. Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. *Trends Genet* 2001;17:589–596.
34. Lerat E, Ochman H. Recognizing the *pseudogenes* in bacterial genomes. *Nucleic Acids Res* 2005;33:3125–3132.
35. Lerat E, Ochman H. Psi-Phi: exploring the outer limits of bacterial *pseudogenes*. *Genome Res* 2004;14:2273–2278.
36. Ortega AP, Villagra NA, Urrutia IM, Valenzuela LM, Talamilla-Espinoza A, *et al*. Lose to win: marT pseudogenization in *Salmonella enterica* serovar Typhi contributed to the surV-dependent survival to H_2O_2, and inside human macrophage-like cells. *Infect Genet Evol* 2016;45:111–121.
37. Sun Y-C, Jarrett CO, Bosio CF, Hinnebusch BJ. Retracing the evolutionary path that led to flea-borne transmission of *Yersinia pestis*. *Cell Host Microbe* 2014;15:578–586.
38. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31:3210–3212.
39. 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.
40. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome: Table 1. *Bioinformatics* 2015;31:2877–2878.
41. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.
42. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* 2018;35:518–522.

43. Moore RM, Harrison AO, McAllister SM, Polson SW, Wommack KE. Iroki: automatic customization and visualization of phylogenetic trees. *PeerJ* 2020;8:e8584.
44. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 2001;29:2607–2618.
45. Kiryutin B, Souvorov A, TatusovaT. ProSplign – protein to genomic alignment tool. 2007.
46. Agarwala R, Barrett T, Beck J, Benson DA, Bollin C. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2018;46:D8–D13.
47. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 2012;3:217–223.
48. 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.
49. 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* 2017.
50. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;16:157.
51. Galens K, Orvis J, Daugherty S, Creasy HH, Angiuoli S, *et al*. The IGS standard operating procedure for automated prokaryotic annotation. *Stand Genomic Sci* 2011;4:244–251.
52. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;47:D309–D314.
53. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000;28:33–36.
54. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, *et al*. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;47:D607–D613.
55. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, *et al*. Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon *Mutagenesis*. *mBio* 2017;8:e02133-16.
56. Gibson AJ, Passmore IJ, Faulkner V, Xia D, Nobeli I, *et al*. Probing differences in gene essentiality between the human and animal adapted lineages of the *Mycobacterium tuberculosis* complex using TnSeq. *Front Vet Sci* 2021;8:760717.
57. Liu BZDJQCLYJ. VFDB 2019: a comparative pathogenomic platform with an interactive web interface | nucleic acids research | oxford academic. *Nucleic Acids Res* 2019;Vol. 47:D687–D692.
58. Malone KM, Rue-Albrecht K, Magee DA, Conlon K, Schubert OT, *et al*. Comparative'omics analyses differentiate *Mycobacterium tuberculosis* and *M*ycobacterium bovis and reveal distinct macrophage responses to infection with the human and bovine tubercle bacilli. *Microb Genom* 2018;4. Epub ahead of print 2018.
59. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;37:907–915.
60. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 2016;44:W3–W10.
61. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;31:166–169.
62. Kelkar DS, Kumar D, Kumar P, Balakrishnan L, Muthusamy B, *et al*. Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. *Mol Cell Proteomics* 2011;10:M111..
63. Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, *et al*. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 2014;32:223–226.
64. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, *et al*. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *PROTEOMICS Clin Appl* 2015;9:745–754.
65. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. *Proteomics* 2013;13:22–24.
66. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;20:1466–1467.
67. Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, *et al*. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* 2011;10:M111..
68. GraphPad, USA LJC. GraphPad Prism version 6 for Windows.
69. Stojković V, Noda-Garcia L, Tawfik DS, Fujimori DG. Antibiotic resistance evolved via inactivation of a ribosomal RNA methylating enzyme. *Nucleic Acids Res* 2016;44:8897–8907.
70. Safi H, Gopal P, Lingaraju S, Ma S, Levine C, *et al*. Phase variation in *Mycobacterium tuberculosis* glpK produces transiently heritable drug tolerance. *Proc Natl Acad Sci* 2019;116:19665–19674.
71. Zumárraga M, Bigi F, Alito A, Romano MI, Cataldi A. A 12.7 kb fragment of the *Mycobacterium tuberculosis* genome is not present in *Mycobacterium bovis*. *Microbiology* 1999;145 (Pt 4):893–897.
72. Gordon SV, Eiglmeier K, Garnier T, Brosch R, Parkhill J, *et al*. Genomics of *Mycobacterium bovis*. *Tuberculosis* 2001;81:157–163.
73. Viana-Niero C, Rodriguez CAR, Bigi F, Zanini MS, Ferreira-Neto JS, *et al*. Identification of an IS6110 insertion site in plcD, the unique phospholipase C gene of *Mycobacterium bovis*. *J Med Microbiol* 2006;55:451–457.
74. Mehta M, Singh A. *Mycobacterium tuberculosis* WhiB3 maintains redox homeostasis and survival in response to reactive oxygen and nitrogen species. *Free Radic Biol Med* 2019;131:50–58.
75. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, *et al*. The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog* 2013;9:e1003543.
76. Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* 2018;16:202–213.
77. Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV. Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* 2009;7:537–544.
78. Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EPC. After the bottleneck: genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Res* 2012;22:721–734.
79. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, *et al*. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol* 2008;6:e311.
80. Lin PL, Ford CB, Coleman MT, Myers AJ, Gawande R, *et al*. Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. *Nat Med* 2014;20:75–79.
81. Dean GS, Rhodes SG, Coad M, Whelan AO, Cockle PJ, *et al*. Minimum infective dose of *Mycobacterium bovis* in cattle. *Infect Immun* 2005;73:6467–6471.
82. Pepperell C, Hoeppner VH, Lipatov M, Wobeser W, Schoolnik GK, *et al*. Bacterial genetic signatures of human social phenomena among *M.tuberculosis* from an aboriginal canadian population. *Mol Biol Evol* 2010;27:427–440.
83. Smith NH, Gordon SV, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG. Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*. *Nat Rev Microbiol* 2006;4:670–681.
84. Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, *et al*. The genome of *Mycobacterium africanum* West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. *PLoS Negl Trop Dis* 2012;6:e1552.

85. Glynn JR, Kremer K, Borgdorff MW, Rodriguez MP, Van Soolingen D. Beijing/W genotype *Mycobacterium tuberculosis* and drug resistance: European concerted action on new generation genetic markers and techniques for the epidemiology and control of tuberculosis. *Emerg Infect Dis* 2006;12:736–743.
86. Parwati I, van Crevel R, van Soolingen D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis* 2010;10:103–111.
87. van der Spuy GD, Kremer K, Ndabambi SL, Beyers N, Dunbar R, *et al*. Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. *Tuberculosis* 2009;89:120–125.
88. Karmakar M, Trauer JM, Ascher DB, Denholm JT. Hyper transmission of Beijing lineage *Mycobacterium tuberculosis*: systematic review and meta-analysis. *J Infect* 2019;79:572–581.
89. Glynn JR, Whiteley J, Bifani PJ, Kremer K, van Soolingen D. Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis:* a systematic review. *Emerg Infect Dis* 2002;8:843–849.
90. Hanekom M, Gey van Pittius NC, McEvoy C, Victor TC, Van Helden PD, *et al*. *Mycobacterium tuberculosis* Beijing genotype: a template for success. *Tuberculosis* 2011;91:510–523.
91. Brosch R, Gordon SV, Pym A, Eiglmeier K, Garnier T, *et al*. Comparative genomics of the mycobacteria. *Int J Med Microbiol* 2000;290:143–152.
92. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, *et al*. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proc Natl Acad Sci* 2015;112:863–868.
93. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, *et al*. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* 2003;35:32–40.
94. Bellerose MM, Baek S-H, Huang C-C, Moss CE, Koh E-I, *et al*. Common variants in the glycerol kinase gene reduce tuberculosis drug efficacy. *mBio* 2019;10:1–15.
95. Gupta A, Alland D. Reversible gene silencing through frameshift indels and frameshift scars provide adaptive plasticity for *Mycobacterium tuberculosis*. *Nat Commun* 2021;12:1–11.
96. Gonzalo-Asensio J, Pérez I, Aguiló N, Uranga S, Picó A, *et al*. New insights into the transposition mechanisms of IS6110 and its dynamic distribution between *Mycobacterium tuberculosis* Complex lineages. *PLoS Genet* 2018;14:1–23.
97. Rosenberg SM. Reverse mutation. In: *In: Brenner's Encyclopedia of Genetics*, Second edition. Elsevier Inc, 2013. pp. 220–221.
98. Mould DL, Stevanovic M, Ashare A, Schultz D, Hogan DA. Metabolic basis for the evolution of a common pathogenic *Pseudomonas aeruginosa* variant. *Elife* 2022;11:e76555.
99. Mehaffy C, Belisle JT, Dobos KM. Mycobacteria and their sweet proteins: an overview of protein glycosylation and lipoglycosylation in *M. tuberculosis*. *Tuberculosis* 2019;115:1–13.
100. Gröschel MI, Sayes F, Simeone R, Majlessi L, Brosch R. ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat Rev Microbiol* 2016;14:677–691.
101. Mishra SK, Shankar U, Jain N, Sikri K, Tyagi JS, *et al*. Characterization of G-quadruplex motifs in espB, espK, and cyp51 genes of *Mycobacterium tuberculosis* as potential drug targets. *Mol Ther Nucleic Acids* 2019;16:698–706.
102. Marjanovic O, Iavarone AT, Riley LW. Sulfolipid accumulation in *Mycobacterium tuberculosis* disrupted in the mce2 operon. *J Microbiol* 2011;49:441–447.
103. Sirakova TD, Thirumala AK, Dubey VS, Sprecher H, Kolattukudy PE. The *Mycobacterium tuberculosis* pks2 gene encodes the synthase for the hepta- and octamethyl-branched fatty acids required for sulfolipid synthesis. *J Biol Chem* 2001;276:16833–16839.
104. Gonzalo Asensio J, Maia C, Ferrer NL, Barilone N, Laval F, *et al*. The virulence-associated two-component PhoP-PhoR system controls the biosynthesis of polyketide-derived lipids in *Mycobacterium tuberculosis*. *J Biol Chem* 2006;281:1313–1316.
105. Chesne-Seck M-L, Barilone N, Boudou F, Gonzalo Asensio J, Kolattukudy PE, *et al*. A point mutation in the two-component regulator PhoP-PhoR accounts for the absence of polyketide-derived acyltrehaloses but not that of phthiocerol dimycocerosates in *Mycobacterium tuberculosis* H37Ra. *J Bacteriol* 2008;190:1329–1334.
106. Gonzalo-Asensio J, Malaga W, Pawlik A, Astarie-Dequeker C, Passemar C, *et al*. Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl Acad Sci* 2014;111:11491–11496.
107. Srinivas M, Rajakumari S, Narayana Y, Joshi B, Katoch VM, *et al*. Functional characterization of the phospholipase C activity of Rv3487c and its localization on the cell wall of *Mycobacterium tuberculosis*. *J Biosci* 2008;33:221–230.
108. Broset E, Martín C, Gonzalo-Asensio J. Evolutionary landscape of the *Mycobacterium tuberculosis* complex from the viewpoint of PhoPR: implications for virulence regulation and application to vaccine development. *mBio* 2015;6:e01289-15.
109. Lee CE, Goodfellow C, Javid-Majd F, Baker EN, Shaun Lott J. The crystal structure of TrpD, a metabolic enzyme essential for lung colonization by *Mycobacterium tuberculosis,* in complex with its substrate phosphoribosylpyrophosphate. *J Mol Biol* 2006;355:784–797.
110. Riley R, Pellegrini M, Eisenberg D. Identifying cognate binding pairs among a large set of paralogs: the case of PE/PPE proteins of *Mycobacterium tuberculosis*. *PLoS Comput Biol* 2008;4:e1000174.
111. Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinform* 2016;17:154–179.
112. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 2016;17:125.
113. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011;13:36–46.
114. Nuccio S-P, Bäumler AJ, Finlay BB. Comparative analysis of *Salmonella* genomes identifies a metabolic network for escalating growth in the inflamed gut. *mBio* 2014;5:e00929–14.
115. Goodhead I, Darby AC. Taking the pseudo out of *pseudogenes*. *Curr Opin Microbiol* 2015;23:102–109.
116. Rambaut A. FigTree v1.4.3. 2012. http://tree.bio.ed.ac.uk/software