Mapping gene-by-gene single-nucleotide variation in 8,535 Mycobacterium tuberculosis genomes

Papakonstantinou, Danai; Dunn, Steven J.; Draper, Simon J.; Cunningham, Adam F.; O'Shea, Matthew K.; McNally, Alan; Achkar, Jacqueline M.

DOI:
10.1128/mSphere.01224-20

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESOURCE REPORT

Mapping Gene-by-Gene Single-Nucleotide Variation in 8,535 *Mycobacterium tuberculosis* Genomes: a Resource To Support Potential Vaccine and Drug Development

Danai Papakonstantinou,[a] Steven J. Dunn,[a] Simon J. Draper,[b] Adam F. Cunningham,[c] Matthew K. O'Shea,[c] Alan McNally[a]

[a]Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom
[b]Jenner Institute, University of Oxford, Oxford, United Kingdom
[c]Institute of Immunology and Immunotherapy, MRC Centre of Immune Regulation, University of Birmingham, Birmingham, United Kingdom

Matthew K. O'Shea and Alan McNally are joint senior authors.

ABSTRACT Tuberculosis (TB) is responsible for millions of deaths annually. More effective vaccines and new antituberculous drugs are essential to control the disease. Numerous genomic studies have advanced our knowledge about *M. tuberculosis* drug resistance, population structure, and transmission patterns. At the same time, reverse vaccinology and drug discovery pipelines have identified potential immunogenic vaccine candidates or drug targets. However, a better understanding of the sequence variation of all the *M. tuberculosis* genes on a large scale could aid in the identification of new vaccine and drug targets. Achieving this was the focus of the current study. Genome sequence data were obtained from online public sources covering seven *M. tuberculosis* lineages. A total of 8,535 genome sequences were mapped against *M. tuberculosis* H37Rv reference genome, in order to identify single nucleotide polymorphisms (SNPs). The results of the initial mapping were further processed, and a frequency distribution of nucleotide variants within genes was identified and further analyzed. The majority of genomic positions in the *M. tuberculosis* H37Rv genome were conserved. Genes with the highest level of conservation were often associated with stress responses and maintenance of redox balance. Conversely, genes with high levels of nucleotide variation were often associated with drug resistance. We have provided a high-resolution analysis of the single-nucleotide variation of all *M. tuberculosis* genes across seven lineages as a resource to support future drug and vaccine development. We have identified a number of highly conserved genes, important in *M. tuberculosis* biology, that could potentially be used as targets for novel vaccine candidates and antituberculous medications.

IMPORTANCE Tuberculosis is an infectious disease caused by the bacterium *Mycobacterium tuberculosis*. In the first half of the 20th century, the discovery of the *Mycobacterium bovis* BCG vaccine and antituberculous drugs heralded a new era in the control of TB. However, combating TB has proven challenging, especially with the emergence of HIV and drug resistance. A major hindrance in TB control is the lack of an effective vaccine, as the efficacy of BCG is geographically variable and provides little protection against pulmonary disease in high-risk groups. Our research is significant because it provides a resource to support future drug and vaccine development. We have achieved this by developing a better understanding of the nucleotide variation of all of the *M. tuberculosis* genes on a large scale and by identifying highly conserved genes that could potentially be used as targets for novel vaccine candidates and antituberculous medications.

Citation Papakonstantinou D, Dunn SJ, Draper SJ, Cunningham AF, O'Shea MK, McNally A. 2021. Mapping gene-by-gene single-nucleotide variation in 8,535 *Mycobacterium tuberculosis* genomes: a resource to support potential vaccine and drug development. mSphere 6:e01224-20. https://doi.org/10.1128/mSphere.01224-20.

Editor Jacqueline M. Achkar, Albert Einstein College of Medicine

Address correspondence to Matthew K. O'Shea, m.k.oshea@bham.ac.uk, or Alan McNally, a.mcnally.1@bham.ac.uk.

Received 1 December 2020
Accepted 10 February 2021
Published 10 March 2021

KEYWORDS *Mycobacterium tuberculosis*, tuberculosis, TB, single nucleotide polymorphisms, SNPs, drug targets, vaccine candidates, single-nucleotide variation, tuberculosis vaccines

Mycobacterium tuberculosis is the causative agent of tuberculosis (TB), an ancient disease that remains one of the biggest causes of infectious disease mortality worldwide (1). According to the WHO, TB accounts for the death of over 1.3 million people every year, while almost 25% of the world's population is latently infected (1). Limitations of diagnostic methods and the emergence of multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB further compound the global health challenge (2).

A major obstacle in combating TB is the lack of an effective vaccine (1, 3). The efficacy of *Mycobacterium bovis* bacillus Calmette-Guérin (BCG), the only commercially available TB vaccine to date, is geographically variable and provides little protection against pulmonary disease (1, 3). At the same time, large-scale genomic analyses have shown that although the *M. tuberculosis* genome exhibits relatively low levels of sequence variation, and undergoes no obvious homologous recombination, strain diversity is more extensive than previously thought (2). The two human *M. tuberculosis* complex species (*M. tuberculosis sensu stricto* and *Mycobacterium africanum*) are composed of seven distinct, globally distributed lineages (L1 to L7), with an eighth lineage recently reported (2, 4).

The spread of *M. tuberculosis* in eight human-adapted lineages, in combination with the complex host immune response to the organism, makes the development of a new globally applicable vaccine or drug more challenging (5). Sequence variation within a drug target-binding site could lead to reduced binding affinity and drug resistance (6, 7). Similarly, gene variation is an important consideration in understanding the likely value of a target antigen in a vaccine, since gene variation is largely responsible for protein variability in the bacterial population (8). In the field of TB, an ideal vaccine would have universal application and an efficacy of >50% against adult pulmonary TB (9). The traditional vaccinology approach has focused on developing TB vaccines based on immunodominant antigens (5). With this approach, only one candidate targeting adults with latent disease has shown protective efficacy of >50%, in phase IIb clinical trials, despite decades of research (5, 10).

Advances in whole-genome sequencing have broadened our knowledge of *M. tuberculosis* in areas such as *M. tuberculosis* population structure (2), transmission patterns, and outbreaks (11) and the evolution of drug resistance (2, 6, 12). At the same time, using reverse vaccinology has led to the identification of potential immunogenic *M. tuberculosis* antigens and T-cell epitopes (13). However, we still do not know what constitutes protective immunity in TB (3). In addition, a large number of *M. tuberculosis* genes and proteins are underexplored for their potential as targets for drug and vaccine strategies. If we are to identify new vaccine and drug targets, then it is important to first identify genes that are highly conserved across the entire species. As such, a better understanding of the sequence variation of every *M. tuberculosis* gene across all lineages could be beneficial for the identification of new proteins and may have potential application in novel drug and vaccine discovery.

We have conducted a large-scale genomic study to provide a high-resolution analysis of the single-nucleotide variation of every *M. tuberculosis* gene across 8,535 genomes. We have combined our analysis with functional annotation from multiple sources. We show that there are extremely high levels of variation in genes known to be involved in drug resistance. Conversely, our approach also shows extreme conservation across the *M. tuberculosis* population in genes involved in latent TB infection and stress responses. These genes are conserved across seven lineages, which suggests an essential role in TB biology, and may constitute novel targets for TB drug and vaccine development.

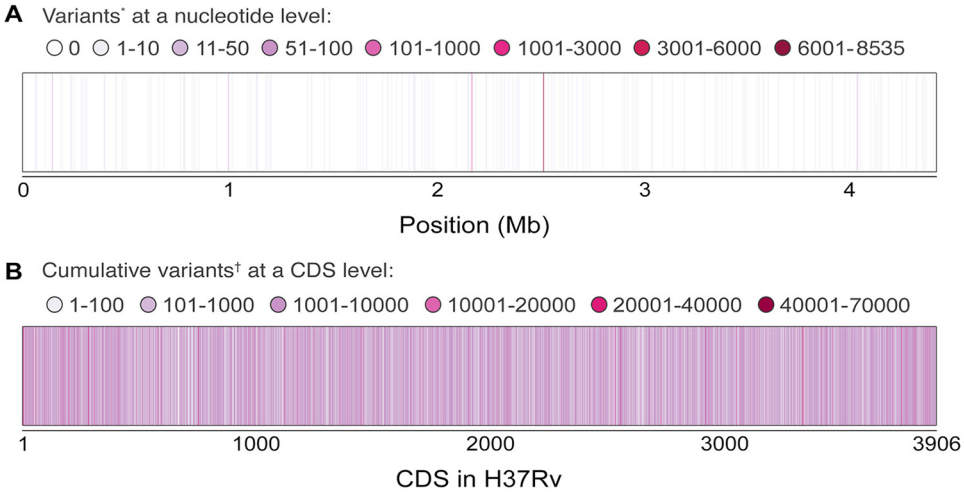

FIG 1 Distribution of the sequence variation at a genomic position and CDS level. (A) Variants*, number of genomes with a variant across H37Rv. Mapping of 8,535 genomes against H37Rv demonstrated that 92.2% of H37Rv genomic positions were conserved upon comparison. A small number of positions contained a large amount of genomes containing a variant (1 to 10 genomes had variants in 7.13% of genomic positions of H37Rv, 10 to 100 genomes had variants in 0.49% of H37Rv genomic positions, and 100 to 8,530 genomes had variants in 0.16% of H37Rv genomic positions. (B) Cumulative variants†, cumulative number of variants across the genomes at a CDS level. CDS from 1 to 3906 and their total number of variants across the data set. At a CDS level, all coding sequences have some degree of variation. Mobile elements, repeat regions, transposases, and RNAs were excluded from this analysis.

RESULTS

Collation of a data set encompassing the known sequenced genomic diversity of *M. tuberculosis*. Our final analysis included 8,535 genome sequences from key genomic study data sets (see Table S1 in the supplemental material). The raw genome data from these studies were mapped against the H37Rv reference genome (14). All seven global lineages were represented in our data, with L1 to L4 being predominant, which is in keeping with sequenced data previously published (12). The breakdown of the lineages in our data was as follows: L4, 56.8%; L2, 21.4%; L3, 12.8%; L1, 7.7%; and L5, L6, and L7, 1.3%. The population structure in our data set is representative of the known sequenced strain diversity of *M. tuberculosis* (2) (Fig. S1).

Nucleotide variation is distributed across the genome within the *M. tuberculosis* population, with a number of variant hot spots. Our analysis revealed that the majority of genomic positions in the *M. tuberculosis* H37Rv genome were completely conserved (92.2%) and that a small number of genomic positions contained variants across sequenced genomes in comparison to that of H37Rv (7.13% of the genomic positions had variants in 1 to 10 genomes, 0.49% had variants in 10 to 100 genomes, and 0.16% had variants in 100 to 8,530 genomes) (Fig. 1A and Fig. S2). However, all identified coding sequences (CDS) in the H37Rv genome had some level of sequence variation across the genomes analyzed (Fig. 1B). Summary statistics of the single nucleotide polymorphism (SNP) distribution at a CDS level showed a mean of 1,403.1 genomes from 8,535 genomes containing a nucleotide variant at a given position in a CDS and a median of 498.5 genomes containing a variant at a given position per gene. To remove potential bias, the results were normalized by gene length, resulting in a mean level of variation of 1.43654 genomes containing a mutation/bp and a median of 0.55407 genomes containing a mutation/bp per CDS.

Genes with high numbers of variants above the mean/median, such as *gyrA*, had hot spots within the gene at which the majority of the mutations occurred. The *gyrA* gene, which is associated with drug resistance to quinolones (6, 15), is 2,517 bp long, and across the analyzed population, the majority of genomic positions within *gyrA* are conserved (87.3%). This gene exhibits variation only at certain positions, some of which are known to be associated with quinolone resistance (e.g., 7,345 genomes had a variant at position 7585).

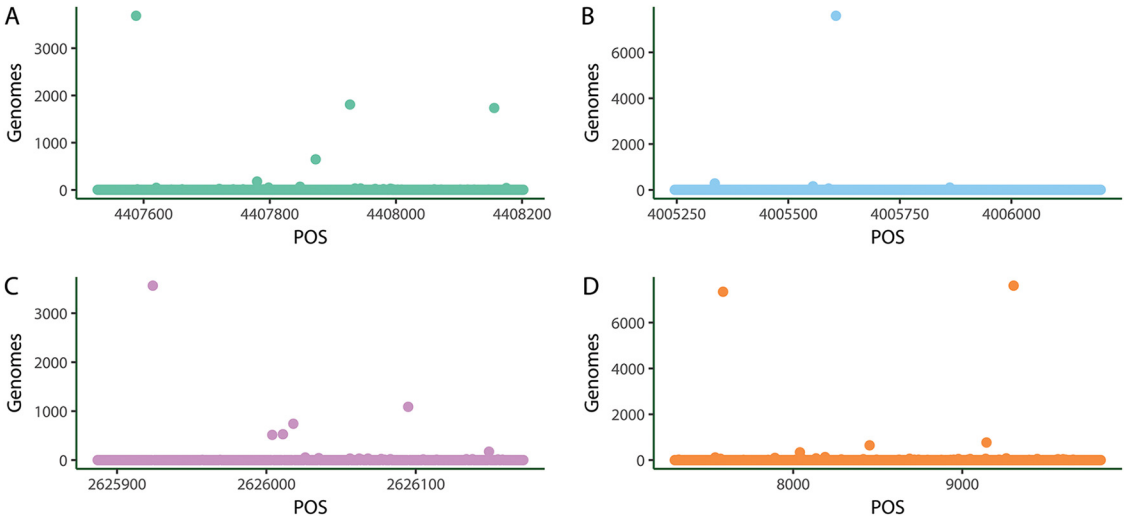

FIG 2 Variant hot spots within *M. tuberculosis* genes *gidB* (A), *fadE33* (B), *esxO* (C), and *gyrA* (D). Four representative examples of genes with high variation: *gidB*, a gene associated with streptomycin resistance (6); *fadE33* (a member of the *fadE* family), which plays a role in cholesterol metabolism (25); *esxO*, which belongs to the ESAT-6 group (16); and *gyrA*, associated with quinolone drug resistance (6). (A) *gidB*. The majority of genomic positions (POS) within *gidB* are conserved. However, 3,690 genomes had a variant at position 4407588, 1,811 genomes had a variant at position 4407927, and 1,734 genomes had a variant at position 4408156. (B) *fadE33*. The majority of genomic positions within *fadE33* are conserved. However, 7,603 genomes had a variant at position 4005607, and 279 genomes had a variant at position 4005335. (C) *esxO*. The majority of genomic positions within *esxO* are conserved. Examples of highly variable genomic areas within *esxO* are positions 2625924 and 2626095. Specifically, 3,561 genomes had a variant at position 2625924 and 1,088 genomes had a variant at position 2626095. (D) *gyrA*. *gyrA*, which is associated with drug resistance to quinolones (6), is 2,517 bp (bp) long, and across the analyzed population, the majority of genomic positions within *gyrA* are conserved. This gene exhibits variation only at certain positions, some of which are known to be associated with quinolone resistance (e.g., 7,345 genomes had a variant at position 7585). In addition, 7,609 and 769 genomes had a variant at genomic positions 9304 and 9143, respectively.

Examples of important families of genes exhibiting high variation (e.g., *gidB* [6], *esxO* [16], *fadE33* [17]), due to such hot spots, are shown in Fig. 2.

Identification of genes with high and low levels of single-nucleotide variation across the data set. We sought to identify genes that were highly conserved and those that contained high levels of sequence variation. We determined a statistical threshold for areas with higher and lower variability than normal. We identified genes in the 5th and 95th percentiles with respect to the amount of sequence variation present in all genomes analyzed (Fig. 3 and Data Set S1). When putative functions were assigned to genes (via clusters of orthologous protein groups [COGs]) in each of the percentiles, the majority of the proteins were uncharacterized or poorly characterized (Table S2). We used Mycobrowser (18) and UniProt (19) databases in an attempt to further characterize genes. The initial comparison of the broader COG categories (e.g., metabolism) did not show any statistically significant difference in gene function between the 5th and 95th percentiles (Table S3). After additional functional annotation using TubercuList categories (14), we noted that more genes related to cell wall and cell processes had high sequence variation (95th percentile, $\chi^2 = 8.8108$, $P < 0.01$) (Table S3). In addition, there were a number of functional groups of proteins that were significantly associated with one of the two percentiles, e.g., toxin-antitoxin (TA) in the 5th percentile and antimicrobial resistance genes in the 95th percentile (Table S3).

We quantified the number of synonymous and nonsynonymous mutations present in the high- and low-variation genes and the percentage of single-nucleotide substitutions for each. CDS in the 95th percentile have more genes with a higher percentage of nonsynonymous SNPs (median, 90.4% of substitutions are nonsynonymous) than genes in the 5th percentile (median, 53% of substitutions are nonsynonymous) (Fig. S3).

Genes with high numbers of SNPs are associated with drug resistance and cell wall-associated processes. Of the 186 genes in the 95th percentile displaying the highest levels of nucleotide variation, 72 contained a functional annotation in H37Rv

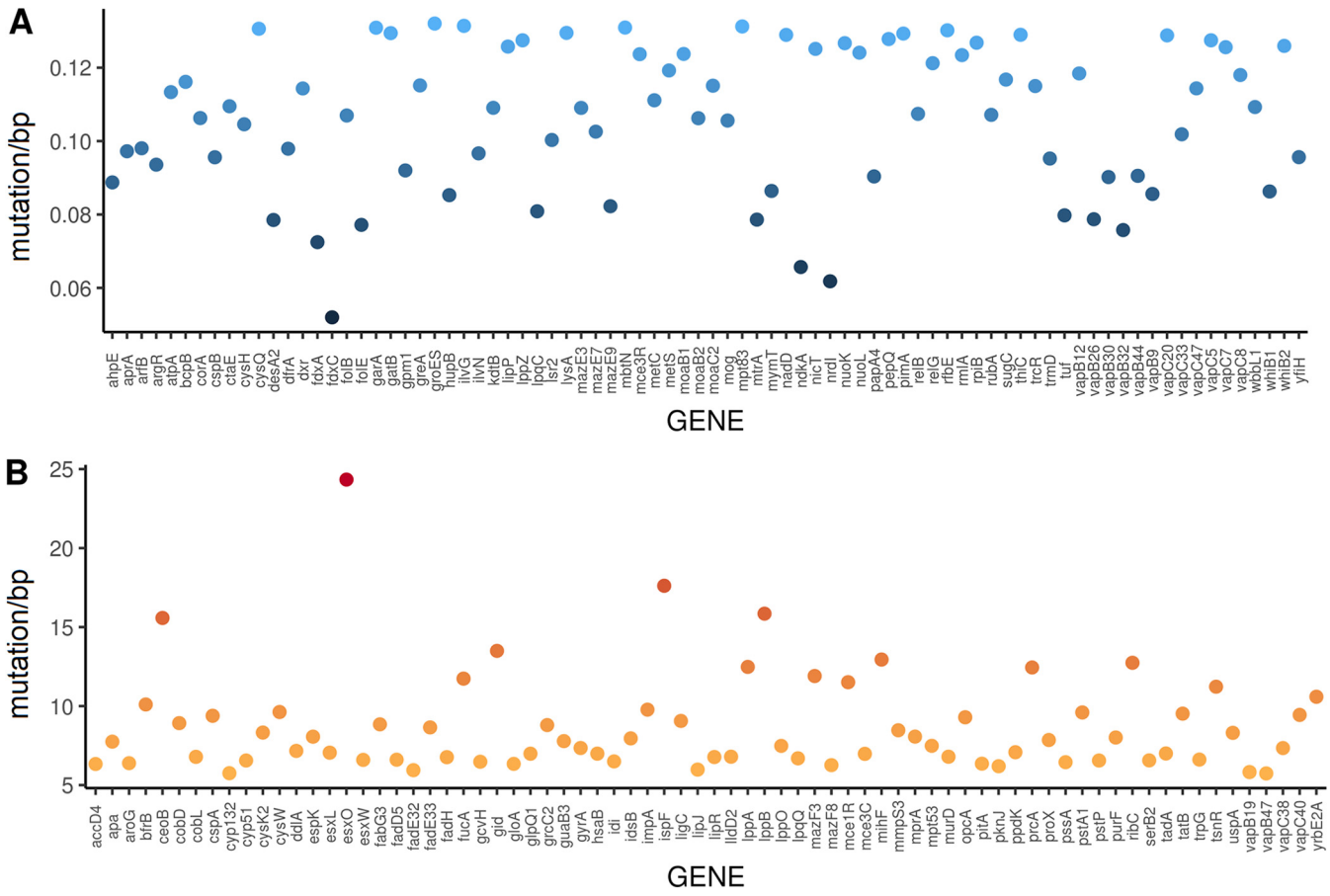

FIG 3 Distribution of the single-nucleotide variation of the genes present in the 5th (A) and 95th (B) percentiles. (A) Fifth percentile genes. Of the genes with functional annotation to H37Rv, the gene with the lowest number of variants across the analyzed genomes in the 5th percentile was *fdxC* (0.05 genomes containing mutations/bp), which encodes a ferredoxin. It was noted that toxin-antitoxin genes of group II are predominant in the 5th percentile. This graph does not contain genes that were not genetically characterized with reference to H37Rv (i.e., hypothetical proteins). Information on these genes can be found in Data Set S1 in the supplemental material. Please note that in order to remove potential bias, the results were normalized by gene length (mutations/bp). (B) Ninety-fifth percentile genes. The *esxO* gene is the first genetically characterized gene with the highest variation (24.3 genomes containing mutations/bp) across the analyzed genomes. Genes associated with drug resistance (e.g., *lppB*, *lppA*, *gidB*), *fadD* and *fadE* families (e.g., *fadE33*, *fadE32*, *fadH*), and ESAT-6 genes are present in the 95th percentile. This graph does not contain genes that were not genetically characterized with reference to H37Rv (i.e., hypothetical proteins). Information on these genes can be found in Data Set S1 in the supplemental material. Please note that in order to remove potential bias, the results were normalized by gene length (mutations/bp).

(Table 1). The remaining 114 genes mainly encoded hypothetical proteins (Data Set S1). Only 14.5% of the genes in the 95th percentile were deemed essential in transposon mutagenesis studies (20).

Functional groups of genes significantly overrepresented in the 95th percentile included drug resistance genes (e.g., *gyrA*) and others related to important *M. tuberculosis* families (e.g., ESAT-6), as well as cell wall and cell processes (e.g., *pitA*, *murD*). A large number of genes demonstrating high sequence variation were previously identified as vaccine (e.g., *esxW*, *apa*) (21) or drug (e.g., *accD4*, *fadE33*, *fadH*, *aroG*, *tatB*, *cyp132*) targets (22) (Fig. 4). Genes associated with drug resistance were also highly variable, including *gidB* (associated with low-level streptomycin resistance) (6), *gyrA* (quinolone-associated resistance gene) (6), and *ceoB* and *opcA* (observed in isoniazid-resistant strains) (23). We also found a high number of genomes with SNPs in *lppA* and *lppB*, which have been identified as novel targets for drug resistance to isoniazid, and in *lldD2*, a novel target for drug resistance to moxifloxacin (24). Genes from the ESAT-6/ESX family (e.g., *esxO*, *espK*), which are associated with virulence and disease pathogenesis (16), were present in the 95th percentile. In fact, *esxW* is considered highly immunogenic and is part of a current vaccine trial (5). Genes from other functional categories related to virulence were also identified (e.g., *tatB*, *mce*) (25). Only nine of the genes in

TABLE 1 Protein prediction in the 95th percentile

Category[a]	COG(s)	Gene(s)	Protein information
Cellular processes and signaling [D],[M],[N],[O],[T],[U],[V],[W],[Y],[Z][b]	D	*vapB47*	Antitoxin, TA group
	M	*gid* (6)	Associated with streptomycin resistance
	M	*murD*	Peptidoglycan biosynthesis
	O	*prcA*[c]	Intermediary metabolism and respiration
	T	*pstP*[c]	Regulatory proteins
	T	*mazF3*	Toxin, TA group
	U	*tatB*[c] (35)	Probable transmembrane transporter
Information storage and processing [A],[B],[J],[K],[L][b]	J	*tsnR, mihF*	Information pathways
	J	*tadA*	CMP-type deaminase domain protein
	K	*mce1R, mprA, pknJ*	Regulatory proteins
	K	*cspA*	Cold shock protein
	L	*gyrA* (6)	Associated with quinolone resistance
	L	*ligC*	DNA recombination and repair
Metabolism [C],[E],[F],[G],[H],[I],[P],[Q][b]	C, H	*fadH,c idsB*	Lipid metabolism
	C, G, Q	*cyp132,*[c] *cyp51,*[c] *ppdK,*[c] *fucA, opcA*	Intermediary metabolism and respiration
	C, E	*aroG,*[c] *glpQ1,*[c] *lldD2,*[c] *mpt53, gcvH*	Intermediary metabolism and respiration
	E	*proX*	Transmembrane transporter activity
	E	*aroG, cysK2, trpG,*[c] *serB2*[c]	Amino acid biosynthesis
	F	*purF*	Purine biosynthesis/purine salvage
	F, P	*ddlA,*[c] *ceoB, uspA, pstA1, pitA*[c]	Cell wall and cell processes
	F, I, G, H	*guaB3,*[c] *lipR, idi, impA*	Intermediate metabolism and respiration
	H	*ribC, cobD, cobL*	Riboflavin/cobalamin biosynthesis
	H, I	*accD4,*[c] *fadD5, fadE32*	Involved in lipid metabolism
	H, I	*fadE33, pssA, fabG3*	Involved in lipid metabolism
	I, H	*lspF,*[c] *grcC2, lipJ*	Intermediate metabolism and respiration
	P	*pstA1, cysW*[c]	Transmembrane transporter activity
	P	*bfrB*	Iron storage protein
	Q	*mce3C* (25)	Virulence factor, Mce family
	Q	*yrbE2A*	Part of mce2 operon
Poorly characterized [R],[S][b]	S	*apa*	Immunogenic, cell wall and cell processes
	S/UC	*hsaB*	Cholesterol catabolism
	S.UC	*vapC38, vapC40, vapB19, mazF8*	TA group
	S/UC	*lppA, lppB, lpqQ, lppO*	Possible lipoproteins
Unable to characterize (UC)	S/UC	*esxO, esxL, esxW*	ESAT-6-like protein
	S/UC	*espK*	ESX-1 secretion system
	S/UC	*mmpS3*	Determinant of intrinsic *M. tuberculosis* AMR

[a]Classification of clusters of orthologous protein groups (COGs) in the 95th percentile, combined with information from Mycobrowser (18) and UniProt (19). Additional information from the literature is individually cited within the table. Genes related to basic COG categories (e.g., metabolism) were observed in both percentiles. However, certain families, such as the *fadD* and *fadE* genes (e.g., *fadE33*, *fadD5*, *fadE32*), associated with fatty acid and cholesterol metabolism, were observed only in the 95th percentile (17). Genes related to pathogenesis of TB disease (e.g., ESAT-6/ESX genes) and antibiotic resistance (e.g., *gyrA*, *gidB*) are present. ESAT-6/ESX family genes were predicted as poorly characterized. The classification-involved proteins encoded by genetically characterized genes with reference to H37Rv. Protein prediction for the noncharacterized genes can be found in Data Set S1 in the supplemental material.

[b]COG subcategories are explained analytically in the legend to Table S2 in the supplemental material.

[c]A number of genes have been identified as high-confidence drug targets (22).

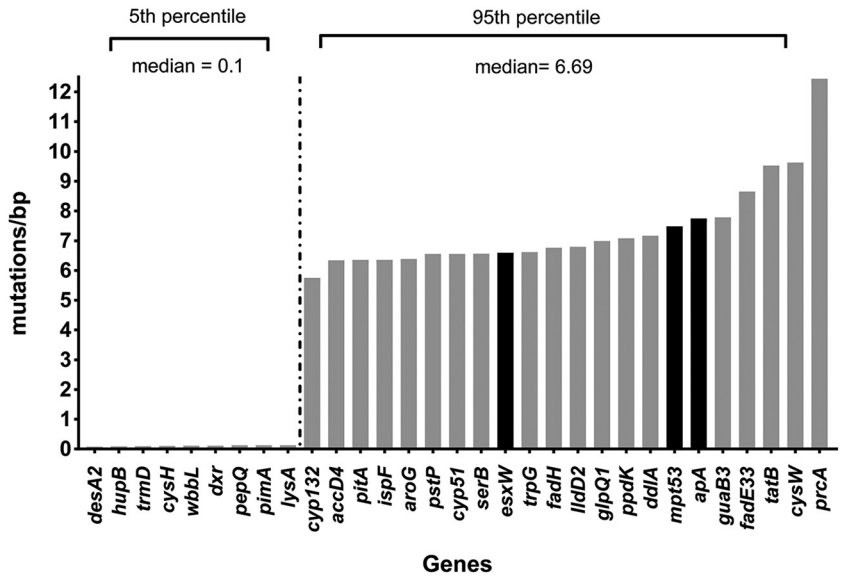

FIG 4 Drug and vaccine candidates previously proposed in the literature within the 5th and 95th percentiles. It is striking that a large number of genes demonstrating high single-nucleotide variation (95th percentile) in our data set have been previously proposed in the literature as desirable drug candidates (gray bars) (22). Few of these drug targets have been previously selected due to their location (e.g., *tatB*) or their biological function (e.g., *fadE33*) (17, 22). In addition, three genes in the 95th percentile have been previously proposed as potential vaccine candidates (*esxW*, *mpt53*, *apa*) (black bars) (5, 21). In fact, *esxW* encodes an immunogenic protein, which is present in a current subunit vaccine (ID93/GLA-SE) (5). A smaller number of genes that are highly conserved in our data set (5th percentile) have been previously proposed as drug targets (gray bars). Genes in the 5th percentile, previously proposed as drug targets, have a median ratio of 0.109272 mutations/bp. Genes in the 95th percentile, previously proposed as drug targets, have a median ratio of 6.691533 mutations/bp. Biological functions and functional annotation of the genes are described in Tables 1 and 2.

the 95th percentile are associated with stress responses (e.g., *cspA*, *bfrB*) (Table S4). Finally, we report that several members of the PE-PPE family (e.g., PPE33, PE_PGRS47, PPE69, PPE18, PE_PGRS4, PPE59, PE_PGRS37, PPE19, PE_PGRS10, PE_PGRS9, PPE57) demonstrated high sequence variation in our initial analysis. However, due to the well-recognized problem of false-positive SNPs identified in these genes during mapping, we excluded them from our subsequent analysis, consistent with the approach of previous studies (26–28) (Data Set S1).

Highly conserved genes are associated mainly with stress responses and maintenance of redox balance. Of a total of 186 genes occurring in the 5th percentile, 82 genes were functionally annotated in the H37Rv genome (14) (Table 2). The remaining 104 genes encoded mainly hypothetical proteins (Data Set S1). In contrast to the genes in the 95th percentile, 28% of the genes in the 5th percentile were deemed essential in transposon mutagenesis studies (20).

A large number of genes that demonstrated low variation were members of the toxin-antitoxin (TA) group of genes. Of the 15 TA genes in the 5th percentile, seven encoded toxins (e.g., *vapC33*) and 10 encoded antitoxins (e.g., *vapB26*). Most of the TA genes with low variation belong to the group II TA system (e.g., *relB*, *mazE9*) (Table 2). TA genes in *M. tuberculosis* have been associated with many processes, including persistence and upregulation under stress conditions (29). Several highly conserved genes were related to redox processes, which can contribute to the protection of *M. tuberculosis* in a hostile environment (e.g., intracellular) (30). Within the 5th percentile, we observed genes encoding metal binding proteins, such as iron-sulfur proteins (*fdxa*, *fdxc*, *whiB1*, *whiB2*) and others encoding proteins that play a role in copper, zinc, or molybdenum metabolism (e.g., *moaB1*, *moaB2*, *mog*, *mymT*) (Table 2). Genes which also demonstrate a redox function or are involved in direct detoxification of peroxides and adaptation of *M. tuberculosis* inside the phagosome are also present in this lower

TABLE 2 Protein prediction in the 5th percentile

Category[a]	COG(s)	Gene(s)	Protein information
Cellular processes and signaling [D],[M],[N],[O],[T], [U], [V],[W],[Y],[Z]	D	*relB*, *relG*	Toxin-antitoxin (TA) group
	M	*rmlA*	Carbohydrate biosynthesis
	M	*ftsQ*	Essential cell division protein
	M	*mpt83*	Cell surface lipoprotein
	O	*groES*	Chaperonin GroES
	O	*ahpE* (64), *bcpB*	Peroxiredoxin (direct antioxidants)
	T	*mtrA*	Response regulator
	T	*garA*	Virulence and glutamate metabolism
	V, M	*lipP*, *pimA*	Role in lipid metabolism
Information storage and processing [A],[B],[J],[K],[L]	J, K, L	*tuf*, *trmD*, *greA*, *hupB*, *gatB*	Information pathways
	K	*mazE9*	Antitoxin (TA group)
	K	*whiB1*, *whiB2*, *mce3R*	Transcriptional regulator
	K	*cspB*	Cold shock protein
	K	*argR*	Amino acid biosynthesis
	K	*trcR*	Regulatory proteins
Metabolism [C],[E],[F],[G],[H],[I],[P],[Q]	C	*fdxA*, *fdxC*	Iron-sulfur proteins
	C	*ctaE*	Probable cytochrome oxidase
	C	*rubA*	Probable rubredoxin
	C, E	*atpA*, *nuoL*, *nuoK*, *pepQ*	Intermediate metabolism and respiration
	E	*metC*, *lysA*, *livN*, *ilvG*	Amino acid transport and metabolism
	F	*nrdI*	Ribonucleotide reductase function
	F, G, H, I	*ndkA*, *rpiB*, *nadD*, *thiC*, *dxR*, *gpm1*	Intermediate metabolism and respiration
	G, Q, I, P	*lppZ*, *lpqC*, *papA4*, *desA2*	Role in lipid metabolism
	G, P	*sugC*, *rfbE*	ABC transporter
	H	*dfrA*, *folE*, *folB*	Involved in folate metabolism
	H	*kdtB*	Coenzyme A (CoA) biosynthesis
	H	*mog*, *moaB2*, *moaB1*, *moaC2* (30)	Molybdopterin biosynthesis
	I	*mbtN*	Mycobactin biosynthesis
	P	*cysH*, *cysQ* (65)	Sulfate activation pathway
	P	*corA*	Transmembrane protein
Poorly characterized [R],[S]	UC	*vapB26*, *vapB44*, *vapB12*, *vapB9*	Antitoxin (TA group)
	S	*vapC33*, *vapC8*, *vapC7*, *vapC20*	Toxin (TA group)
	S	*vapC5*, *vapC47*, *vapB30*, *vapB32*	TA group
	S	*yfiH*	Multicopper oxidase
Unable to characterize (UC)	UC, S	*wbbL1*, *nicT*, *arfB*	Cell and cell wall-associated processes
	UC	*lsr2*	Nucleoid-associated protein Lsr2
	UC	*mazE7*, *mazE3*	Antitoxin (TA group)
	UC	*mymT* (37)	Metallothionein
	UC	*aprA* (66)	Acid and phagosome regulated protein
	UC	*metS*	Information pathways

[a]COG classification of proteins combined with information from Mycobrowser (18) and UniProt (19). Additional information from the literature, which cannot be found in these two databases, is individually cited within the table. Genes related to the metabolism of essential elements for *M. tuberculosis* survival, such as thiamine (e.g., *thiC*), and others related to cell envelope and active transport were also observed (e.g., *sugC*). Genes belonging to the TA family, as well as genes related to metal binding and antioxidant activity, are present in the 5th percentile. The majority of the TA genes are poorly characterized by COGs. The classification-involved proteins were encoded by genetically characterized genes with reference to H37Rv. Protein prediction for the noncharacterized genes can be found in Data Set S1 in the supplemental material.

part of the SNP distribution. Examples of these groups of genes are those encoding peroxiredoxins (*ahpE*, *bcpB*), rubredoxins (*rubA*), or proteins associated with the sulfur activation pathway (*cysQ*, *cysH*). Many of the genes mentioned above have been previously found to be upregulated or downregulated under stress conditions in order to possibly protect the bacillus from a hostile environment (e.g., inside macrophage) (30). We found 34 genes that have been previously identified to be influenced under stress (Table S5). Only one gene (*dfrA*) associated with drug resistance to isoniazid and *para*-aminosalicylic acid (PAS) (6, 31) was present in the 5th percentile, but its role as a resistance target has been previously debated (31).

DISCUSSION

Large-scale genomic studies have revolutionized our knowledge of *M. tuberculosis* drug resistance, population structure, and outbreak transmission patterns (2, 6, 11, 12). At the same time, multiple new antigens have been identified as novel drug and

vaccine targets by numerous methods, including reverse vaccinology (13). A deeper understanding of the sequence variation present across all *M. tuberculosis* genes (and their resultant products) on a large scale could identify biologically important genes which are highly conserved. These, in turn, could be used as potential targets for novel vaccines or therapeutics. Thus, it is important to examine the levels of sequence variation in target genes, as this can directly relate to drug resistance or vaccine escape. We have combined functional annotation from multiple sources into a large-scale genomic study that provides a "snapshot" of the single-nucleotide variation of every *M. tuberculosis* gene and that could be used as a tool to identify desirable novel vaccine candidates and antituberculous drugs.

From our analysis, we have observed that the genes with the highest levels of single-nucleotide variation (95th percentile) included many associated with drug resistance (e.g., *gyrA*, *gidB*), pathogenesis (e.g., important gene families such as ESAT-6/ESX), and cell wall biology (e.g., *tatB*) (16, 22, 25). For example, ESAT-6 is a family of 23 proteins secreted by the ESX-1 to ESX-5 secretory proteins in pathogenic *M. tuberculosis* strains. Their potential role in virulence and pathogenesis and their immunogenic nature have made them promising vaccine candidates. In our study, we observed some members of the ESAT-6 and ESX-1 family with high sequence variation (95th percentile). However, any conclusions regarding these genes' variation should be examined with caution due to the known issues of short read mapping in these repeat-rich genes (16, 32).

It is striking that a high number of genes observed in the 95th percentile have been previously identified as drug targets (22). For example, the *fadE* and *fadD* families, as well as genes such as *tatB* and *pitA*, have been previously proposed as desirable drug targets due to their functional properties and location within the organism (17, 22). *fadE* genes are associated with cholesterol metabolism, whereas *fadD* genes are involved in fatty acid metabolism. Both families are known to play a role in important biological processes, including virulence and disease pathogenesis (17, 33, 34). Similarly, genes such as *tatB* and *pitA* are both membrane transporters that have attracted the interest of the TB community as potential drug targets (35). Given the high variation of these genes in our distribution (i.e., 95th percentile), the choice of these gene products as potential drug targets should be considered with caution, as even a few single-nucleotide substitutions can have a significant impact on the binding of a drug (6). For example, if a protein possesses the desirable criteria as a drug target but its gene demonstrates high variation, perhaps several other parameters should be taken into consideration, such as the position of this variation, whether this position is in a function-critical site (e.g., active binding site), and the nature of the SNPs affected (synonymous versus nonsynonymous). Considering these points, it is important to minimize the possibility that large swaths of the *M. tuberculosis* population could easily become resistant to any targeting agent.

Similarly, in the 95th percentile, we have also observed genes encoding immunogenic proteins and membrane proteins with high variation that have been identified as potential vaccine candidates (e.g., *mpt53*) (21). A protein that is well recognized by adaptive immunity may not constitute a desirable vaccine target due to epitope variation in the bacterial population. If such a protein is part of a vaccine, then SNPs within the gene encoding this protein may influence the level of protection afforded after immunization (8). Moreover, this nucleotide variation may be reflected differently for targets of T-cell responses, where epitopes are linear peptides, than for targets of B-cell responses, where most epitopes are discontinuous and formed by the spatial arrangement of different regions of a protein. Nevertheless, a protein can contain multiple epitopes, and therefore SNPs may affect some responses more than others, as was recently shown for *Salmonella* (8). While it has been shown that T-cell epitopes in *M. tuberculosis* are generally conserved, there are exceptions to this (36). Therefore, assessing the sequence variation of each gene and the genomic positions where this variation most frequently occurs may inform future vaccine design.

Our study can be used as an overall assessment of the SNPs of any *M. tuberculosis* gene across a large data set. More importantly, this study provides a list of more than 80 genetically characterized genes with very low nucleotide variation in the sequenced population, the majority of which are associated with latent TB infection, survival in stress environments, and maintenance of redox balance. These genes are conserved across seven lineages, which suggests an essential role in the TB infection, and may constitute novel targets for TB drug and vaccine development. In this group of highly conserved genes (5th percentile), there are characteristic examples with interesting biological function, such as genes involved in the sulfur assimilation pathway and others associated with metal binding and metal metabolism (30, 37). Overall, these genes play a role in the maintenance of redox balance during *M. tuberculosis* infection (30, 38). For instance, we have observed two 7Fe ferredoxin genes (*fdxC* and *fdxA*) (39) with very low sequence variation and interesting biological function that may constitute interesting drug or vaccine targets. FdxA is a ferredoxin that has been found to be expressed under hypoxic and acidic conditions, resembling the hostile environment inside the macrophage during infection (39). FdxA is also part of the DosR regulon, a group of proteins that play a critical role during anaerobic metabolism and latent TB (40, 41). Notably, the DosR regulon is involved in latent infection, when the bacillus needs "protection" from the hostile environment inside the macrophage, and its proteins are activated by hypoxia, nitric oxide (NO), and carbon monoxide (CO) (42, 43). The DosR regulon has attracted significant interest, with vaccine candidates based on DosR proteins currently under evaluation (42, 43). Similarly, *fdxC*, which encodes another 7Fe ferredoxin, demonstrated the lowest sequence variation across our data set and has been identified as an essential gene in transposon libraries (20, 39). FdxC has not been previously synthesized; however, the crystal structure of FdxA (which shares great genetic similarity with FdxC) has already been established and can be used as a model for future FdxC synthesis (39). Redox metabolic pathways and maintenance of redox homeostasis have already been the center of interest in the development of novel *M. tuberculosis* drug targets. Other genes demonstrating low variation (i.e., 5th percentile) that are involved in redox homeostasis or adaptation within the macrophage are genes related to molybdenum biosynthesis (e.g., *moaB1* and *moaB2*) and others involved in sulfur metabolism (such as *cysQ*). The molybdenum metabolism plays a very important role in nitrate respiration and the ability of *M. tuberculosis* to persist in lung granulomas under hypoxic stress conditions (30, 44). Sulfur metabolites are part of the sulfur assimilation pathway and are essential for the *M. tuberculosis* adaptation and survival inside the macrophage. Key genes of the sulfur assimilation pathway seem to be induced under differential stress conditions which resemble the latent stage of TB infection. For example, CysQ is a 3′-phosphoadenosine-5′-phosphatase with an important regulatory role in *M. tuberculosis* sulfur pathway and has been already expressed in an *Escherichia coli* vector and induced under nutrient starvation (45). The conservation of some of these genes, their interesting biological properties, and the fact that they are often missing a human equivalent make them attractive drug candidates.

There are certain limitations of this study. We used H37Rv as a reference, which is a common strain used for genomic studies. H37Rv is a lineage 4 strain, known to have unique sequence variation in comparison to other strains (26, 46). We have accounted for this bias by masking all the genomic positions where more than 8,050 strains (95% of the data set) contained an identical nucleotide variant relative to H37Rv. To confirm that this masking did not lead us to miss variations present in masked sites within the population, we looked at the allelic profile of every H37Rv coordinate in our data set. The total number of multiallelic sites in our data set represents only 0.19% of the H37Rv genome (i.e., 8,220 sites), the vast majority of which are minor frequency events (median number of isolates with a variant at a multiallelic site, 3). Only 18 multiallelic sites (~0.000408% of the entire genome) have major frequency events (i.e., a nucleotide that is present in the majority of the data set). We also acknowledge that there is

differential gene content between other lineages and H37Rv (26, 47). However, the aim of this study is to specifically identify genes core to all *M. tuberculosis* lineages that can be used as common drug targets or vaccine candidates. In this study, we have used a large number of published genomes, which gives us a cross-representation of susceptible and resistant strains, i.e., what is currently circulating in TB. We acknowledge that we have utilized a number of drug resistance studies (see Table S1 in the supplemental material), as they are readily available online. This is probably a reflection of the great interest of the TB community in drug resistance during the last decade (6, 12, 24, 48). While the aim of this study was not to define drug resistance mutations, we show that genomic positions previously associated with drug resistance demonstrate high variation in our data set (e.g., positions 7585 and 9304 in *gyrA* or positions 4407588 and 4407927 in *gidB*) (24, 49, 50). Finally, although we appreciate the important role of mobile elements or PE-PPE genes in the *M. tuberculosis* genome, we have excluded them from our final analysis (along with other repetitive regions) due to the increased chance of false-positive SNPs in these regions. For instance, the PE-PPE family of genes are responsible for more than 10% of the coding capacity in the *M. tuberculosis* genome and are considered to play an important role in pathogenesis, antigenic variation, virulence, and immune modulation (51–53). There is also increasing interest in the scientific community in PE-PPE genes as novel drug and vaccine candidates. However, PE-PPE genes are typically excluded from SNP calling and phylogenetic reconstruction because of false-positive variants during mapping, derived from issues of mapping short-read sequencing against genes with multiple repeat regions (26, 28, 54).

In summary, we present a gene-by-gene analysis of single-nucleotide variation across a well-curated selection of 8,535 *M. tuberculosis* genomes as a resource to support future drug and vaccine development. We identified a number of highly conserved genes which could potentially be considered targets for novel vaccine candidates and for antituberculous medications.

MATERIALS AND METHODS

Sequence data and initial filtering. We assembled a collection of genomes by utilizing large-scale genome data sets from a number of key *M. tuberculosis* genomic projects (see Table S1 in the supplemental material). Sequence data of 8,931 *M. tuberculosis* clinical isolates were downloaded from the public domain (https://www.ebi.ac.uk/ena) in the form of paired-end fastq files. The project accession numbers are available in Table S1. Reads were trimmed using Sickle (v1.33) (55) configured to a quality threshold of 20 over a 50-bp sliding window. The average sequencing depth of all genomes was determined. The final data set has a mean coverage of 103.5 with a standard deviation of 73.4.

Variant calling and mapping. The filtered reads were processed with Snippy (v3.2 Dev) (56) and were mapped against the reference genome for *M. tuberculosis* H37Rv (54). Snippy was configured to require a minimum base quality of 20, a minimum read coverage of 10, and a minimum allele frequency of 0.9 in order to obtain high-confidence variants. Snippy-core was used to determine the mapping coverage of each sequence, and a cutoff of 90% mapping coverage was chosen to filter the data set. This specific cutoff was selected after mapping three *Mycobacterium canettii* strains (which shares a common ancestor with *M. tuberculosis*) against H37Rv and determining the mapping coverage. All sequences below 90% mapping coverage were discarded. Following the application of all the quality control measures mentioned in the above sections, the data set contained a net total of 8,535 sequences.

The output from Snippy was concatenated and manually processed in order to identify variants on both a per-sample and a data set-wide basis. Further processing was performed using a script in R (https://github.com/Danaipap/TB_project). The tab-delimited output files derived from each distinct genome's Snippy analysis (i.e., snps.tab) were combined and processed in R in order to produce a single file, which contained information regarding global genome positions (positions 1 to 4411532, as defined in H37Rv), the total number of variants at those positions, and the corresponding gene and protein names. The H37Rv reference genome contains multiple overlapping areas within coding positions. When a variant occurs in one of these areas, Snippy v3.2 Dev (using SNPef) reports the mutation in only one of the coding regions, not the other. This could potentially result in underestimating the number of variants in some overlapping genomic areas. Using the bedtools (57) intersect function, we "intersected" the reference GenBank file with a file containing the exact number of H37Rv genomic positions and the genome ID. Subsequently, we combined this intersected reference file with our combined output file. We excluded noncoding regions, insertions-deletions (indels), and RNAs from our subsequent analysis, as the focus of this study was to assess the overall nucleotide variation in genes core to all *M.*

tuberculosis lineages that can be used as common drug targets or vaccine candidates. We also excluded mobile elements, repeat regions, transposases, and all the PE-PPE family genes from the subsequent analysis, accounting for the known caveat of false-positive SNPs in these genes during mapping (26, 28).

To account for the differential presence or absence of any given gene in the data set, we included only genes that were present at a depth of 10 and greater (minimum required for mapping) across 95% of all tested isolates. To account for the possibility that many variants are unique to H37Rv and so appear with high variation in our analysis (i.e., all genomes are identical but different from the H37Rv genome), we masked all positions where more than 8,050 strains (95% of the data set) contained an identical nucleotide variant relative to H37Rv. This extra validation was performed since H37Rv is a lineage 4 strain and it is known that it carries a lot of sequence variation not present in any other *M. tuberculosis* strain (26). A file describing variants at each nucleotide position in H37Rv (prior to masking and exclusion of certain sites such as those for PE-PPE, etc.) is available at https://github.com/Danaipap/TB_project.

Determining a statistical threshold for areas with higher and lower variability than normal. We normalized our results by dividing the number of mutations in a given gene by the gene's length. It was noted that the distribution of genes with single-nucleotide polymorphisms (SNPs) across the 8,535 mapped genome sequences was not normal. We introduced a systematic way to analytically quantify "low" and "high" levels of sequenced variation. We defined this notion through the rate of change according to which the distribution increases and chose a cutoff where the distribution becomes stable.

We looked at the sequence of the values in each percentile. We call this sequence $f < f_0 = 0.004553734$, $f_1 = 0.085642255\ldots$ $f_{100} = 55.166666667 >$ (see Data Set S2 in the supplemental material). We used the ratio $\rho_i = f_{i+1}/f_i$ to determine the cutoff point where stabilization occurs in the sequence. Since the ratio is >1, the discrete sequence f fits as a mild exponential function. We determined the low-variability segment of the sequence as the points where the function has stabilized to attain the approximate form of c^n, for a constant c. To achieve this, we devised a method that is governed by two parameters: (i) δ, which is the allowed fluctuation $|\rho_{i+1} - \rho_i| \leq \delta$, and (ii) L, the window length (consecutive values) inside which the fluctuation is at most δ. We chose a δ of 0.03 and an L of 3. That is, we looked to satisfy $|\rho_{i+1} - \rho_i| \leq 0.03$ for three consecutive values. For this choice of parameters, we identified as the first stabilization point 5% of the distribution (mutations/bp). Therefore, we chose this position where the stabilization occurs to be our cutoff for this study. Symmetrically, we chose the other side of the distribution (95%), which is typical for statistical studies.

Phylogenetic and population structure analysis. We used raxml with rapid bootstrapping (58) (100 bootstrap replicates) to reconstruct a core SNP phylogeny, using a general GTR+CAT approximation algorithm. The likelihood of the final tree was evaluated and optimized under GAMMA. The multifasta alignment of 8,535 genome sequences was created using the snippy-core function in Snippy. All the PE-PPE regions were masked, and the produced alignment was subsequently cleaned with the snippy-clean function. A core SNP alignment was extracted using the program snp-sites (v2.4.0) (59). One *M. canettii* genome was used as an outgroup. The phylogeny was visualized and processed further in Itol (v3) (60). Additional population structure analysis was performed with fastBAPS (61) (v1.0.0) (hierarchical Bayesian statistical clustering) to determine specific clusters within our sequenced data and to aid with the annotation of the phylogenetic tree at https://github.com/Danaipap/TB_project.

Functional annotation and comparison of group of genes in the 5th and 95th percentiles. Functional annotation of loci in the 5th and 95th percentiles of the diversity distribution was performed with EggNOG-mapper (v4.5) (62), and clusters of orthologous protein groups (COGs) were assigned. Individual loci were extracted using GBKsplit (63) (https://github.com/stevenjdunn/gbkSPLIT). UniProt (19) and Mycobrowser (18) databases were used in order to complement the functional annotation of the proteins. A chi-square test with Yates' continuity correction for the comparison of the frequency of gene groups in the 5th and 95th percentiles was performed in R, and a P value of <0.05 was deemed significant.

Data availability. Accession numbers for the reads used in this project along with the information regarding the year and place of isolation are listed in Table S1 in the supplemental material.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

DATA SET S1, XLSX file, 0.1 MB.
DATA SET S2, XLSX file, 0.3 MB.
FIG S1, PDF file, 2.7 MB.
FIG S2, PDF file, 0.8 MB.
FIG S3, PDF file, 1.5 MB.
TABLE S1, PDF file, 0.2 MB.
TABLE S2, PDF file, 0.2 MB.
TABLE S3, PDF file, 0.02 MB.
TABLE S4, PDF file, 0.1 MB.
TABLE S5, PDF file, 0.3 MB.

ACKNOWLEDGMENTS

We thank Periklis A. Papakonstantinou for devising the protocol of determining a statistical threshold for areas of low and high variation. We thank Chris Connor and Emily Richardson for help and advice with bioinformatics analysis and Spyros Paraskeyas for help at the initial stages of this work with respect to the programming environment and data mining techniques. We thank Rad Poplawski for the outstanding computational technical support and Cloud Infrastructure for Microbial Bioinformatics (CLIMB). We thank Marisol Perez-Toledo and Erin Aldera for their ongoing help and support.

This project was funded by a Global Challenges Ph.D. studentship awarded to D.P. by the University of Birmingham. S.J.D. (Dunn) is funded by BBSRC grant no. BB/R006261/1, awarded to A.M. S.J.D. (Draper) is a Wellcome Trust Senior Fellow (106917/Z/15/Z).

A.M. and M.K.O. conceived and designed the project (corresponding authors). D.P. and S.J.D. (Dunn) analyzed all the data. D.P., S.J.D. (Dunn), A.M., M.K.O., and A.F.C. interpreted the data. All authors wrote and corrected the paper.

We declare that there are no competing interests.

REFERENCES

1. WHO. 2018. Global tuberculosis report 2018. World Health Organization, Geneva, Switzerland.
2. Gagneux S. 2018. Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol 16:202–213. https://doi.org/10.1038/nrmicro.2018.8.
3. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MPR. 2013. The immune response in tuberculosis. Annu Rev Immunol 31:475–527. https://doi.org/10.1146/annurev-immunol-032712-095939.
4. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, Antoine R, Niyigena EB, Mulders W, Fissette K, Diels M, Gaudin C, Duthoy S, Ssengooba W, André E, Kaswa MK, Habimana YM, Brites D, Affolabi D, Mazarati JB, de Jong BC, Rigouts L, Gagneux S, Meehan CJ, Supply P. 2020. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. Nat Commun 11:2917. https://doi.org/10.1038/s41467-020-16626-6.
5. Andersen P, Scriba TJ. 2019. Moving tuberculosis vaccines from theory to practice. Nat Rev Immunol 19:550–562. https://doi.org/10.1038/s41577-019-0174-z.
6. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, Abdallah AM, Alghamdi S, Alsomali M, Ahmed AO, Portelli S, Oppong Y, Alves A, Bessa TB, Campino S, Caws M, Chatterjee A, Crampin AC, Dheda K, Furnham N, Glynn JR, Grandjean L, Minh Ha D, Hasan R, Hasan Z, Hibberd ML, Joloba M, Jones-López EC, Matsumoto T, Miranda A, Moore DJ, Mocillo N, Panaiotov S, Parkhill J, Penha C, Perdigão J, Portugal I, Rchiad Z, Robledo J, Sheen P, Shesha NT, Sirgel FA, Sola C, Oliveira Sousa E, Streicher EM, Van Helden P, Viveiros M, Warren RM, McNerney R, Pain A, . 2018. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet 50:307–316. https://doi.org/10.1038/s41588-017-0029-0.
7. Guo C, D'Ippolito AM, Reddy TE. 2015. Leading edge previews from prescription to transcription: genome sequence as drug target. Cell 162:16–17. https://doi.org/10.1016/j.cell.2015.06.033.
8. Domínguez-Medina CC, Pérez-Toledo M, Schager AE, Marshall JL, Cook CN, Bobat S, Hwang H, Chun BJ, Logan E, Bryant JA, Channell WM, Morris FC, Jossi SE, Alshayea A, Rossiter AE, Barrow PA, Horsnell WG, MacLennan CA, Henderson IR, Lakey JH, Gumbart JC, López-Macías C, Bavro VN, Cunningham AF. 2020. Outer membrane protein size and LPS O-antigen define protective antibody targeting to the Salmonella surface. Nat Commun 11:851. https://doi.org/10.1038/s41467-020-14655-9.
9. Initiative for Vaccine Research, WHO. 2018. WHO preferred product characteristics for new tuberculosis vaccines. Department of Immunization, Vaccines and Biologicals, World Health Organization, Geneva, Switzerland.
10. Tait DR, Hatherill M, Van Der Meeren O, Ginsberg AM, Van Brakel E, Salaun B, Scriba TJ, Akite EJ, Ayles HM, Bollaerts A, Demoitié M-A, Diacon A, Evans TG, Gillard P, Hellström E, Innes JC, Lempicki M, Malahleha M, Martinson N, Mesia Vela D, Muyoyeta M, Nduba V, Pascal TG, Tameris M, Thienemann F, Wilkinson RJ, Roman F. 2019. Final analysis of a trial of M72/AS01$_E$ vaccine to prevent tuberculosis. N Engl J Med 381:2429–2439. https://doi.org/10.1056/NEJMoa1909953.
11. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TEA. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study. Lancet Infect Dis 13:137–146. https://doi.org/10.1016/S1473-3099(12)70277-3.
12. Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE, Brand J, Chapman SB, Cho S-N, Gabrielian A, Gomez J, Jodals AM, Joloba M, Jureen P, Lee JS, Malinga L, Maiga M, Nordenberg D, Noroc E, Romancenco E, Salazar A, Ssengooba W, Velayati AA, Winglee K, Zalutskaya A, Via LE, Cassell GH, Dorman SE, Ellner J, Farnia P, Galagan JE, Rosenthal A, Crudu V, Homorodean D, Hsueh P-R, Narayanan S, Pym AS, Skrahina A, Swaminathan S, Van der Walt M, Alland D, Bishai WR, Cohen T, Hoffner S, Birren BW, Earl AM, Narayanan S, Pym AS, Skrahina A, Swaminathan S, TBResist Global Genome Consortium, . 2017. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. Nat Genet 49:395–402. https://doi.org/10.1038/ng.3767.
13. Tian Y, Da Silva Antunes R, Sidney J, Arlehamn CSL, Grifoni A, Dhanda SK, Paul S, Peters B, Weiskopf D, Sette A. 2018. A review on T cell epitopes identified using prediction and cell-mediated immune models for mycobacterium tuberculosis and bordetella pertussis. Front Immunol 9:2778. https://doi.org/10.3389/fimmu.2018.02778.
14. Lew JM, Kapopoulou A, Jones LM, Cole ST. 2011. TubercuList—10 years after. Tuberculosis (Edinb) 91:1–7. https://doi.org/10.1016/j.tube.2010.09.008.
15. Chien J-Y, Chiu W-Y, Chien S-T, Chiang C-J, Yu C-J, Hsueh P-R. 2016. Mutations in gyrA and gyrB among fluoroquinolone- and multidrug-resistant Mycobacterium tuberculosis isolates. Antimicrob Agents Chemother 60:2090–2096. https://doi.org/10.1128/AAC.01049-15.
16. Brodin P, Rosenkrands I, Andersen P, Cole ST, Brosch R. 2004. ESAT-6 proteins: protective antigens and virulence factors? Trends Microbiol 12:500–508. https://doi.org/10.1016/j.tim.2004.09.007.
17. Wipperman MF, Yang M, Thomas ST, Sampson NS. 2013. Shrinking the fadE proteome of mycobacterium tuberculosis: insights into cholesterol metabolism through identification of an $\alpha 2\beta 2$ heterotetrameric acyl coenzyme a dehydrogenase family. J Bacteriol 195:4331–4341. https://doi.org/10.1128/JB.00502-13.
18. Kapopoulou A, Lew JM, Cole ST. 2011. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis (Edinb) 91:8–13. https://doi.org/10.1016/j.tube.2010.09.006.
19. The UniProt Consortium. 2018. UniProt: the universal protein knowledgebase. Nucleic Acids Res 46:2699. https://doi.org/10.1093/nar/gky092.
20. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, Rubin EJ, Schnappinger D, Ehrt S, Fortune SM, Sassetti CM, Ioerger TR. 2017. Comprehensive essentiality analysis of the Mycobacterium tuberculosis genome via saturating transposon mutagenesis. mBio 8:e02133-16. https://doi.org/10.1128/mBio.02133-16.

21. Wang L, Liu Z, Wang J, Liu H, Wu J, Tang T, Li H, Yang H, Qin L, Ma D, Chen J, Liu F, Wang P, Zheng R, Song P, Zhou Y, Cui Z, Wu X, Huang X, Liang H, Zhang S, Cao J, Wu C, Chen Y, Su D, Chen X, Zeng G, Ge B. 2019. Oxidization of TGFβ-activated kinase by MPT53 is required for immunity to Mycobacterium tuberculosis. Nat Microbiol 4:1378–1388. https://doi.org/10.1038/s41564-019-0436-3.
22. Raman K, Yeturu K, Chandra N. 2008. targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC Syst Biol 2:109. https://doi.org/10.1186/1752-0509-2-109.
23. Jiang X, Zhang W, Gao F, Huang Y, Lv C, Wang H. 2006. Comparison of the proteome of isoniazid-resistant and -susceptible strains of *Mycobacterium tuberculosis*. Microb Drug Resist 12:231–238. https://doi.org/10.1089/mdr.2006.12.231.
24. Mortimer TD, Weber AM, Pepperell CS, Gilbert JA. 2018. Signatures of selection at drug resistance loci in Mycobacterium tuberculosis. mSystems 3:e00108-17. https://doi.org/10.1128/mSystems.00108-17.
25. Forrellad MA, Klepp LI, Gioffré A, Sabio y García J, Morbidoni HR, de la Paz Santangelo M, Cataldi AA, Bigi F. 2013. Virulence factors of the Mycobacterium tuberculosis complex. Virulence 4:3–66. https://doi.org/10.4161/viru.22329.
26. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR, Guthrie JL, Laukens K, Miotto P, Ofori-Anyinam B, Dreyer V, Supply P, Suresh A, Utpatel C, van Soolingen D, Zhou Y, Ashton PM, Brites D, Cabibbe AM, Jong BC, de Vos M, Menardo F, Gagneux S, Gao Q, Heupink TH, Liu Q, Loiseau C, Rigouts L, Rodwell TC, Tagliani E, Walker TM, Warren RM, Zhao Y, Zignol M, Schito M, Gardy J, Cirillo DM, Niemann S, Comas I, Van Rie A. 2019. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. Nat Rev Microbiol 17:533–545. https://doi.org/10.1038/s41579-019-0214-5.
27. Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro BJ, Mcintosh F, Soualhine H, Menzies D, Behr MA, Nathan CF. 2015. Population genomics of Mycobacterium tuberculosis in the Inuit. Proc Natl Acad Sci U S A 112:13609–13614. https://doi.org/10.1073/pnas.1507071112.
28. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S. 2013. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med 10:e1001387. https://doi.org/10.1371/journal.pmed.1001387.
29. Ramage HR, Connolly LE, Cox JS. 2009. Comprehensive functional analysis of Mycobacterium tuberculosis toxin-antitoxin systems: implications for pathogenesis, stress responses, and evolution. PLoS Genet 5:e1000767. https://doi.org/10.1371/journal.pgen.1000767.
30. Levillain F, Poquet Y, Mallet L, Mazères S, Marceau M, Brosch R, Bange F-C, Supply P, Magalon A, Neyrolles O. 2017. Horizontal acquisition of a hypoxia-responsive molybdenum cofactor biosynthesis pathway contributed to Mycobacterium tuberculosis pathoadaptation. PLoS Pathog 13:e1006752. https://doi.org/10.1371/journal.ppat.1006752.
31. Wang F, Jain P, Gulten G, Liu Z, Feng Y, Ganesula K, Motiwala AS, Ioerger TR, Alland D, Vilchèze C, Jacobs WR, Sacchettini JC. 2010. Mycobacterium tuberculosis dihydrofolate reductase is not a target relevant to the antitubercular activity of isoniazid. Antimicrob Agents Chemother 54:3776–3782. https://doi.org/10.1128/AAC.00453-10.
32. Uplekar S, Heym B, Friocourt V, Rougemont J, Cole ST. 2011. Comparative genomics of Esx genes from clinical isolates of Mycobacterium tuberculosis provides evidence for gene conversion and epitope variation. Infect Immun 79:4042–4049. https://doi.org/10.1128/IAI.05344-11.
33. Wipperman MF, Sampson NS, Thomas ST. 2014. Pathogen roid rage: cholesterol utilization by Mycobacterium tuberculosis. Crit Rev Biochem Mol Biol 49:269–293. https://doi.org/10.3109/10409238.2014.895700.
34. Liu Z, Ioerger TR, Wang F, Sacchettini JC. 2013. Structures of mycobacterium tuberculosis FadD10 protein reveal a new type of adenylate-forming enzyme. J Biol Chem 288:18473–18483. https://doi.org/10.1074/jbc.M113.466912.
35. Feltcher ME, Sullivan JT, Braunstein M. 2010. Protein export systems of Mycobacterium tuberculosis: novel targets for drug development? Future Microbiol 5:1581–1597. https://doi.org/10.2217/fmb.10.112.
36. Ernst JD. 2017. Antigenic variation and immune escape in the MTBC, p 171–190. In Advances in experimental medicine and biology. Springer, New York, NY.
37. Gold B, Deng H, Bryk R, Vargas D, Eliezer D, Roberts J, Jiang X, Nathan C. 2008. Identification of a copper-binding metallothionein in pathogenic mycobacteria. Nat Chem Biol 4:609–616. https://doi.org/10.1038/nchembio.109.
38. Chawla M, Parikh P, Saxena A, Munshi M, Mehta M, Mai D, Srivastava AK, Narasimhulu KV, Redding KE, Vashi N, Kumar D, Steyn AJC, Singh A. 2012. Mycobacterium tuberculosis WhiB4 regulates oxidative stress response to modulate survival and dissemination in vivo. Mol Microbiol 85:1148–1165. https://doi.org/10.1111/j.1365-2958.2012.08165.x.
39. Ricagno S, de Rosa M, Aliverti A, Zanetti G, Bolognesi M. 2007. The crystal structure of FdxA, a 7Fe ferredoxin from Mycobacterium smegmatis. Biochem Biophys Res Commun 360:97–102. https://doi.org/10.1016/j.bbrc.2007.06.013.
40. Chen T, He L, Deng W, Xie J. 2013. The Mycobacterium DosR regulon structure and diversity revealed by comparative genomic analysis. J Cell Biochem 114:1–6. https://doi.org/10.1002/jcb.24302.
41. Sherman DR, Voskuil M, Schnappinger D, Liao R, Harrell MI, Schoolnik GK. 2001. Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding-crystallin. Proc Natl Acad Sci U S A 98:7534–7539. https://doi.org/10.1073/pnas.121172498.
42. Voskuil MI, Schlesinger LS. 2015. Toward resolving the paradox of the critical role of the DosR regulon in *Mycobacterium tuberculosis* persistence and active disease. Am J Respir Crit Care Med 191:1103–1105. https://doi.org/10.1164/rccm.201503-0424ED.
43. Mehra S, Foreman TW, Didier PJ, Ahsan MH, Hudock TA, Kissee R, Golden NA, Gautam US, Johnson A-M, Alvarez X, Russell-Lodrigue KE, Doyle LA, Roy CJ, Niu T, Blanchard JL, Khader SA, Lackner AA, Sherman DR, Kaushal D. 2015. The DosR regulon modulates adaptive immunity and is essential for Mycobacterium tuberculosis persistence. Am J Respir Crit Care Med 191:1185–1196. https://doi.org/10.1164/rccm.201408-1502OC.
44. Kaufholdt D, Baillie C-K, Meinen R, Mendel RR, Hänsch R. 2017. The molybdenum cofactor biosynthesis network: *in vivo* protein-protein interactions of an actin associated multi-protein complex. Front Plant Sci 8:1946. https://doi.org/10.3389/fpls.2017.01946.
45. Hatzios SK, Iavarone AT, Bertozzi CR. 2008. Rv2131c from Mycobacterium tuberculosis Is a CysQ 3′-phosphoadenosine-5′-phosphatase. Biochemistry 47:5823–5831. https://doi.org/10.1021/bi702453s.
46. Ioerger TR, Feng Y, Ganesula K, Chen X, Dobos KM, Fortune S, Jacobs WR, Mizrahi V, Parish T, Rubin E, Sassetti C, Sacchettini JC. 2010. Variation among genome sequences of H37Rv strains of Mycobacterium tuberculosis from multiple laboratories. J Bacteriol 192:3645–3653. https://doi.org/10.1128/JB.00166-10.
47. Periwal V, Patowary A, Vellarikkal SK, Gupta A, Singh M, Mittal A, Jeyapaul S, Chauhan RK, Singh AV, Singh PK, Garg P, Katoch VM, Katoch K, Chauhan DS, Sivasubbu S, Scaria V. 2015. Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of Mycobacterium tuberculosis pangenome. PLoS One 10:e0122979. https://doi.org/10.1371/journal.pone.0122979.
48. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T, Drobniewski F. 2014. Evolution and transmission of drug resistant tuberculosis in a Russian population. Nat Genet 46:279–286. https://doi.org/10.1038/ng.2878.
49. Regmi SM, Coker OO, Kulawonganunchai S, Tongsima S, Prammananan T, Viratyosin W, Thaipisuttikul I, Chaiprasert A. 2015. Polymorphisms in drug-resistant-related genes shared among drug-resistant and pan-susceptible strains of sequence type 10, Beijing family of Mycobacterium tuberculosis. Int J Mycobacteriol 4:67–72. https://doi.org/10.1016/j.ijmyco.2014.11.050.
50. Coker OO, Chaiprasert A, Ngamphiw C, Tongsima S, Regmi SM, Clark TG, Ong RTH, Teo YY, Prammananan T, Palittapongarnpim P. 2016. Genetic signatures of Mycobacterium tuberculosis Nonthaburi genotype revealed by whole genome analysis of isolates from tuberculous meningitis patients in Thailand. PeerJ 4:e1905. https://doi.org/10.7717/peerj.1905.
51. Akhter Y, Ehebauer MT, Mukhopadhyay S, Hasnain SE. 2012. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: Perhaps more? Biochimie 94:110–116. https://doi.org/10.1016/j.biochi.2011.09.026.
52. Brennan MJ. 2017. The enigmatic PE/PPE multigene family of mycobacteria and tuberculosis vaccination. Infect Immun 85:e00969-16. https://doi.org/10.1128/IAI.00969-16.
53. Meena LS. 2015. An overview to understand the role of PE_PGRS family proteins in *Mycobacterium tuberculosis* H_{37} Rv and their potential as new drug targets. Biotechnol Appl Biochem 62:145–153. https://doi.org/10.1002/bab.1266.
54. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S,

Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393:537–544. https://doi.org/10.1038/31159.

55. Joshi NF, Fass JN. 2011. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33). (Software.) https://github.com/najoshi/sickle.
56. Seeman T. 2015. Snippy—rapid bacterial SNP calling and core genome alignments. (version 3.2.0). (Software.) https://github.com/tseemann/snippy.
57. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.
58. Stamatakis A, Hoover P, Rougemont J, Renner S. 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst Biol 57:758–771. https://doi.org/10.1080/10635150802429642.
59. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2:e000056. https://doi.org/10.1099/mgen.0.000056.
60. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W244. https://doi.org/10.1093/nar/gkw290.
61. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. 2019. Fast hierarchical Bayesian analysis of population structure. Nucleic Acids Res 47:5539–5549. https://doi.org/10.1093/nar/gkz361.
62. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293. https://doi.org/10.1093/nar/gkv1248.
63. Dunn S. 2018. GBKSplit: extracts genes using a list of locus tags from gbk to logically named nucleotide fasta's. (Software.) https://github.com/stevenjdunn/gbkSPLIT.
64. Hugo MM, Turell L, Manta B, Botti H, Monteiro G, Luis Netto ES, Alvarez B, Radi R, Trujillo M. 2009. Thiol and sulfenic acid oxidation of AhpE, the one-cysteine peroxiredoxin from Mycobacterium tuberculosis: kinetics, acidity constants, and conformational dynamics. Biochemistry 48:9416–9426. https://doi.org/10.1021/bi901221s.
65. Hatzios SK, Bertozzi CR. 2011. The regulation of sulfur metabolism in Mycobacterium tuberculosis. PLoS Pathog 7:e1002036. https://doi.org/10.1371/journal.ppat.1002036.
66. Abramovitch RB, Rohde KH, Hsu F-F, Russell DG. 2011. aprABC: a Mycobacterium tuberculosis complex-specific locus that modulates pH-driven adaptation to the macrophage phagosome. Mol Microbiol 80:678–694. https://doi.org/10.1111/j.1365-2958.2011.07601.x.