

RESEARCH ARTICLE

# Whole Genome Sequencing of *Mycobacterium africanum* Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction

Kathryn Winglee<sup>1</sup>✉, Abigail Manson McGuire<sup>2</sup>✉, Mamoudou Maiga<sup>1,3</sup>✉, Thomas Abeel<sup>2,4</sup>, Terrance Shea<sup>2</sup>, Christopher A. Desjardins<sup>2</sup>, Bassirou Diarra<sup>3</sup>, Bocar Baya<sup>3</sup>, Moumine Sanogo<sup>3</sup>, Souleymane Diallo<sup>3</sup>, Ashlee M. Earl<sup>2\*</sup>, William R. Bishai<sup>1\*</sup>

**1** Center for Tuberculosis Research, Department of Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America, **2** Genome Sequencing and Analysis Program, The Broad Institute of MIT & Harvard, Cambridge, Massachusetts, United States of America, **3** Project SEREFO (Centre de Recherche et de Formation sur le VIH/Sida et la Tuberculose)/University of Sciences, Technics and Technologies of Bamako (USTTB), Bamako, Mali, **4** Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

✉ These authors contributed equally to this work.

✉ Current address: National Institute of Health, Bethesda, Maryland, United States of America

\* [aearyl@broadinstitute.org](mailto:aearyl@broadinstitute.org) (AME); [wbishai1@jhmi.edu](mailto:wbishai1@jhmi.edu) (WRB)



CrossMark  
click for updates

**OPEN ACCESS**

**Citation:** Winglee K, Manson McGuire A, Maiga M, Abeel T, Shea T, Desjardins CA, et al. (2016) Whole Genome Sequencing of *Mycobacterium africanum* Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction. PLoS Negl Trop Dis 10(1): e0004332. doi:10.1371/journal.pntd.0004332

**Editor:** Pamela L. C. Small, University of Tennessee, UNITED STATES

**Received:** June 15, 2015

**Accepted:** December 5, 2015

**Published:** January 11, 2016

**Copyright:** © 2016 Winglee et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All sequencing data are available from GenBank (accession numbers JLAZ000000000-JJVF000000000; see [S1 Table](#)).

**Funding:** This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services, under Contract No.:HHSN272200900018C and Grant Number U19AI110818 to the Broad Institute. In addition, this work has been funded by the Howard Hughes Medical Institute and the National Institutes

## Abstract

### Background

*Mycobacterium africanum*, made up of lineages 5 and 6 within the *Mycobacterium tuberculosis* complex (MTC), causes up to half of all tuberculosis cases in West Africa, but is rarely found outside of this region. The reasons for this geographical restriction remain unknown. Possible reasons include a geographically restricted animal reservoir, a unique preference for hosts of West African ethnicity, and an inability to compete with other lineages outside of West Africa. These latter two hypotheses could be caused by loss of fitness or altered interactions with the host immune system.

### Methodology/Principal Findings

We sequenced 92 MTC clinical isolates from Mali, including two lineage 5 and 24 lineage 6 strains. Our genome sequencing assembly, alignment, phylogeny and average nucleotide identity analyses enabled us to identify features that typify lineages 5 and 6 and made clear that these lineages do not constitute a distinct species within the MTC. We found that in Mali, lineage 6 and lineage 4 strains have similar levels of diversity and evolve drug resistance through similar mechanisms. In the process, we identified a putative novel streptomycin resistance mutation. In addition, we found evidence of person-to-person transmission of lineage 6 isolates and showed that lineage 6 is not enriched for mutations in virulence-associated genes.

of Health grants AI037856, AI036973 and AI097138 to WRB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Conclusions

This is the largest collection of lineage 5 and 6 whole genome sequences to date, and our assembly and alignment data provide valuable insights into what distinguishes these lineages from other MTC lineages. Lineages 5 and 6 do not appear to be geographically restricted due to an inability to transmit between West African hosts or to an elevated number of mutations in virulence-associated genes. However, lineage-specific mutations, such as mutations in cell wall structure, secretion systems and cofactor biosynthesis, provide alternative mechanisms that may lead to host specificity.

## Author Summary

*Mycobacterium africanum* consists of two lineages, lineages 5 and 6, within the *Mycobacterium tuberculosis* complex (MTC) that cause human tuberculosis in West Africa, but are found rarely outside of this region. Our analysis of the whole genome sequences of 26 lineage 5 and 6 isolates, and 66 isolates from other lineages within the MTC, reveal that *M. africanum* does not meet modern criteria to be considered an independent species. We analyzed drug resistance-associated genes and found that drug resistance evolves within these lineages through similar mechanisms as observed for the rest of the MTC in Mali. Though we did not see an elevated number of mutations in virulence-associated genes in these two lineages, we identified a number of lineage-specific mutations, pseudogenes and changes in gene content that may impact virulence and host specificity, and improve, overall, our understanding of what make these lineages unique.

## Introduction

*Mycobacterium africanum* is a member of the *Mycobacterium tuberculosis* complex (MTC) that causes up to half of all tuberculosis cases in West Africa [1]. First identified by Castets in 1968, it was originally characterized as having biochemical characteristics intermediate between *Mycobacterium tuberculosis*, which consists of lineages 1, 2, 3, 4, and 7 and is the main cause of human tuberculosis, and *Mycobacterium bovis*, an animal-adapted lineage that causes bovine tuberculosis [2]. Later work divided *M. africanum* into two lineages, *M. africanum* West African type I and *M. africanum* West African type II, which became known as lineages 5 and 6, respectively, within the MTC [3, 4].

Lineages 5 and 6 cause a disease similar to classically defined *M. tuberculosis*, although it has been suggested that human disease caused by these lineages may differ compared to that caused by lineages 1–4. For example, patients with lineage 6 disease have been reported to show attenuated ESAT-6 responses compared to patients with classical *M. tuberculosis* lineage disease [5, 6]. In addition, in liquid culture systems it has been reported that *M. africanum* has a slower growth rate with a larger bacillary size than *M. tuberculosis* [7, 8]. While some studies have found that *M. africanum* is less virulent than *M. tuberculosis*, both in animal models and human patients [7, 9–11], others show that there is no difference [12]. Though these contradicting results may be due to differences in the study populations, they underscore how little is known about lineages 5 and 6.

Contributing to this lack of knowledge, while lineages 1–4 are widely distributed around the globe, lineages 5–7 are limited to certain regions of Africa [13]. Lineage 7 has only been found

in Ethiopia [14], and lineages 5 and 6 are found almost exclusively in patients living in West Africa, with very few cases occurring outside of this region, mostly involving recent immigrants from West Africa [1]. The reason for the apparent geographic restriction of lineages 5 and 6 is unknown. One hypothesis is the presence of an undiscovered animal reservoir endemic to West Africa, which is supported by the close relationship between lineages 5 and 6 and the animal-adapted lineages of the MTC [15, 16]. Another hypothesis is that lineages 5 and 6 have a unique predilection for humans with genetic backgrounds common in West Africa. In fact, using a retrospective epidemiological study of the MTC in Ghana, Asante-Poku et al. showed that lineage 5 is associated with the Ewe ethnic group [17]. A third hypothesis is that lineages 5 and 6 are unable to compete with other lineages outside of West Africa, either due to loss of fitness or decreased transmissibility, thus explaining their limited global distribution [7].

Historically, mycobacterial subspecies were defined by biochemical assays, but, as genetic tools became more readily available, it is now possible to identify genomic regions that define MTC lineages [18]. The publication of the whole genome sequence of *M. africanum* GM041182, a single lineage 6 strain, provided valuable insights into the genetics of this lineage [19]. For instance, the authors identified lineage 6-specific pseudogenes, a novel region not present in *M. tuberculosis*, and single nucleotide polymorphisms (SNPs) in key genes, all of which may play a role in the geographic restriction of lineage 6. A later study sequenced four additional lineage 6 isolates and was able to confirm many of these findings, but also showed that not all mutations identified in *M. africanum* GM041182 are shared by other members of this lineage [8]. To our knowledge, no study has closely analyzed the genetics of lineage 5.

From these studies, it is clear that more sequenced isolates are needed to fully characterize the genetics of lineages 5 and 6 and to illuminate mechanisms that may explain its geographic isolation. Toward this end, we sequenced 92 clinical MTC isolates from Mali, a country in West Africa in which 26.2% and 1.6% of tuberculosis cases are caused by lineage 6 and lineage 5, respectively [20] [1]. Using these and previously published data, we performed both alignment- and assembly-based comparative analyses to further refine our understanding of lineage-specific genomic features that might explain the geographic distribution of lineages 5 and 6. To our knowledge, this is the largest collection of lineage 6 strains sequenced to date, and the first in depth whole genomic characterization of lineage 5.

## Materials and Methods

### Samples

101 strains were selected from clinical isolates collected in Bamako, Mali [20], and included all strains identified by spoligotyping as *M. africanum*, *M. tuberculosis* T1, or *M. bovis*. Of these strains, 92 were still viable and were submitted for whole genome sequencing. These 92 strains will be referred to as the “Mali Collection” (S1A Table). In addition, to improve MTC lineage representation, we selected additional whole genome assemblies that matched the quality of our assemblies. These included four finished *M. bovis* genomes available from GenBank (*M. bovis* AF2122/97 [21], *M. bovis* BCG Mexico [22], *M. bovis* BCG Pasteur 1173P2 [23], and *M. bovis* BCG Tokyo 172 [24]), a set of 40 *M. tuberculosis* strains (9 lineage 1 strains, 12 lineage 2 strains, 7 lineage 3 strains, and 12 lineage 4 strains) from South Africa [25], the finished *M. africanum* genome from Genbank (*M. africanum* GM041182) [19], and our outgroup, *M. canettii* CIPT 140010059 [26]. Combined with the Mali Collection, these 137 strains will be referred to as the “Assembly Collection” (S1B Table). Finally, all 161 strains (122 lineage 2, two lineage 3, and 37 lineage 4) from a study in China were included in the variant analysis to improve geographical and lineage representation [27]. The samples from the China study (S1C

[Table](#)) combined with the samples from South Africa and Mali (for a total of 289 strains) will be referred to as the “Alignment Collection”.

## Ethics statement

The study protocols for the Mali samples were approved by the Ethics Committee of the University of Bamako and the Institutional Review Board of the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIAID/NIH), Bethesda, MD, USA. For all samples, written informed consent was obtained from study participants prior to cohort enrollment [20]. For the South African samples, Biomedical Research Ethics Council (BREC) approval from the University of KwaZulu-Natal was granted for collection of sputum specimens from study participants and for whole genome sequencing of clinical strains. Written informed consent was obtained from study participants prior to cohort enrollment, or waived by BREC [25].

## Drug susceptibility testing

Drug resistance to isoniazid, rifampicin, ethambutol and streptomycin was tested for all Mali strains as previously described [20]. We confirmed those results by submitting 17 strains to National Jewish Health in Colorado for agar proportion testing of isoniazid, rifampicin, ethambutol, ofloxacin, niacin, kanamycin, ethionamide, capreomycin, amikacin, cycloserine and para-aminosalicylic acid, as well as radiometric testing of ciprofloxacin and pyrazinamide. The agar proportion results confirmed the mycobacterial growth indicator tube (MGIT) tests performed in Mali. Genotypic drug resistance was determined for rifampicin, isoniazid, ethambutol, streptomycin, ofloxacin, kanamycin and ethionamide using genetic markers from line-probe assays (S2 Table).

## Genome sequencing

Extraction of genomic DNA was performed on 10 mL cultures grown in 7H9 broth using the CTAB-lysozyme method as previously described [28]. Library preparation and whole genome sequencing (WGS) were performed as previously described [29–31]. GenBank accessions for all strains used in this analysis can be found in S1B Table, along with assembly statistics for the new sequences generated at the Broad Institute (92 sequences from Mali generated for this study, and 40 sequences from South Africa).

## Annotation

All genomes in our Assembly Collection were uniformly annotated by transferring annotations from *M. tuberculosis* H37Rv. The reference *M. tuberculosis* H37Rv genome (accession CP003248.2) was aligned to draft assemblies using Nucmer [32]. This alignment was used to map reference genes over to the target genomes. Using this methodology, annotations were successfully transferred onto all 137 strains for 3466 of the *M. tuberculosis* H37Rv genes; the rest of the *M. tuberculosis* H37Rv genes transferred to a subset of the genomes.

For those genes not cleanly mapping to *M. tuberculosis* H37Rv, the protein-coding genes were predicted with the software tool Prodigal [33]. tRNAs were identified by tRNAscan-SE [34] and rRNA genes were predicted using RNAmmer [35]. Gene product names were assigned based on top blast hits against the SwissProt protein database ( $> = 70\%$  identity and  $> = 70\%$  query coverage), and protein family profile search against the TIGRfam hmmer equivalents. Additional annotation analyses performed include Pfam [36], TIGRfam [37], Kyoto Encyclopedia of Genes and Genomes (KEGG) [38], clusters of orthologous groups

(COG) [39], Gene Ontology (GO) [40], enzyme commission (EC) [41], SignalP [42], and Transmembrane Helices; Hidden Markov Model (TMHMM) [43].

### Digital spoligotyping

Reads from each isolate were aligned against the 43 spacer sequences traditionally used in wet lab spoligotyping [28, 44]. From these alignments, the number of matching reads was used to determine if the spacer was present. The spacer was considered absent if the read count total was in the lowest quartile of counts. Spacers were defined as present by using a Bonferroni corrected p-value based on an exponential distribution of the average absent spacer counts. If the p-value was <0.01 the spacer was considered to be present. The spacer pattern was matched to the SITVITWEB database to generate a named spoligotype for each isolate and to determine the spoligotype international type (SIT) [45].

### Orthogroup clustering and phylogenetic trees

SYNERGY2 [46–48], available at <http://sourceforge.net/projects/synergytwo/>, was used to identify cluster-based orthogroups across our Assembly Collection of 137 genomes, which we will refer to as “SYNERGY orthogroups”. In addition, for each *M. tuberculosis* H37Rv gene, we defined a second set of annotation transfer-based ortholog groups as the set of genes for which annotations were transferred from this *M. tuberculosis* H37Rv gene in our annotation protocol, which we will refer to as “*M. tuberculosis* H37Rv-based orthologs”. Genes without *M. tuberculosis* H37Rv orthologs were manually examined in the context of their SYNERGY orthogroups to identify lineage-specific novel genes.

Phylogenetic trees were generated by applying RAxML [49] to a concatenated alignment of 3343 single-copy core SYNERGY orthogroups (excluding orthogroups with paralogs) across all 137 organisms. Bootstrapping was performed using RAxML’s rapid bootstrapping algorithm (1000 iterations).

### Average nucleotide identity analysis (ANI)

Calculations of ANI were done as previously described [50, 51] using the SYNERGY orthogroups calculated from the Assembly Collection.

### Gene content analysis

PAUP [52] was used to reconstruct gain and loss of *M. tuberculosis* H37Rv-based orthologs at ancestral nodes of the Assembly Collection phylogenetic tree using parsimony. In order to analyze changes in gene content, we used a cost matrix with values of 10 for a gene gain, 5 for a gene loss, and 0.2 for an increase or decrease in copy number. We looked for orthologs found within all members of one clade, and absent in other clades. As a further filter, we also required that orthogroups be found in >80% of the clade of interest, and <20% of other strains. We performed this analysis for four key clades: lineage 5, lineage 6, the clade including *M. bovis* and lineage 6, and the clade including lineages 5, 6 and *M. bovis*.

In addition, we selected the Pfam gene categories most expanded or reduced in each clade of interest. We determined significance using Fisher’s test ( $Q < 0.05$ ). For each of the clades described above, we compared the strains below this node versus all other strains in our analysis.

## Identification of SNPs

For our Alignment Collection, reads were mapped onto a reference strain of *M. tuberculosis* H37Rv (GenBank accession number CP003248.2) using BWA version 0.5.9.9 [53]. In cases where read coverage of the reference was greater than 200x, reads were down-sampled using Picard [54] prior to mapping. Variants, including both single nucleotide polymorphisms (SNPs) and multi-nucleotide polymorphisms, were identified using Pilon version 1.5 as described [29] and were used to construct phylogenetic trees using FastTree [55].

We defined lineage-specific variants for lineage X, as those occurring in at least 95% of the strains of lineage X (true positive rate >95%), missing in less than 5% of strains of lineage X (positive predictive value >95%), occurring in less than 5% of the strains that do not belong to lineage X (true negative rate >95%) and not occurring in at least 95% of strains not belonging to lineage X (negative predictive value >95%). The absolute number of true positives must exceed seven. Formulas are schematically presented in S3 Table.

Mutations were considered *M. africanum*-specific (lineage 5 and 6-specific, identified as LIN-Maf in S4 and S5 Tables) if they met these cutoffs for lineage 5 and 6 combined, and were present in both lineage 5 strains. Similarly, mutations were considered *M. tuberculosis*-specific if they met these cutoffs for lineages 1–4 combined. No *M. tuberculosis*-specific mutations were identified. Due to inclusion of only two lineage 5 strains in our dataset, no lineage-specific variants were identified in lineage 5. Thus, for this lineage only, we used a less stringent requirement to define lineage-specific variants: we required that variants be present in both lineage 5 strains and in <5% of the strains in each other lineage. We classified each gene containing a lineage-specific variant into functional group categories, including GO [40], KEGG [38], Pfam [36], and COG [39]. We then evaluated enrichment using Fisher's Exact test and corrected for multiple comparisons using the Storey method for functional group categories [56].

## Identification of pseudogenes

A pseudogene was defined as any gene that had a loss of function mutation anywhere within the coding sequence. Loss of function mutations were defined as nonsense mutations, or insertions or deletions with lengths that were not multiples of 3 base pairs or were greater than 30 base pairs. Lineage-specific pseudogenes were determined using the same definitions as for variants on a per gene basis (positive predictive value > 95%, negative predictive value >95%, true positive rate >95%, true negative rate >95%, number of true positives >7, with the exception of lineage 5, which used the SNP cutoffs of pseudogene in both lineage 5 strains and in <5% in each other lineage).

## Computational gene function assessments

The effect of select non-synonymous mutations on protein function was assessed using the online version of SIFT at default settings [57], unless there was low confidence in the prediction, in which case SIFT was run for each of the four available databases (UniRef90 from April 2011 [default], UniProt-SwissProt 57.15 from April 2011, UniProt-TrEMBL from March 2009 and NCBI nonredundant from March 2011). Peptide binding was predicted using the NetMHC-II online tool with default settings [58].

## Results

### *M. africanum* and *M. tuberculosis* lineages are part of the same species

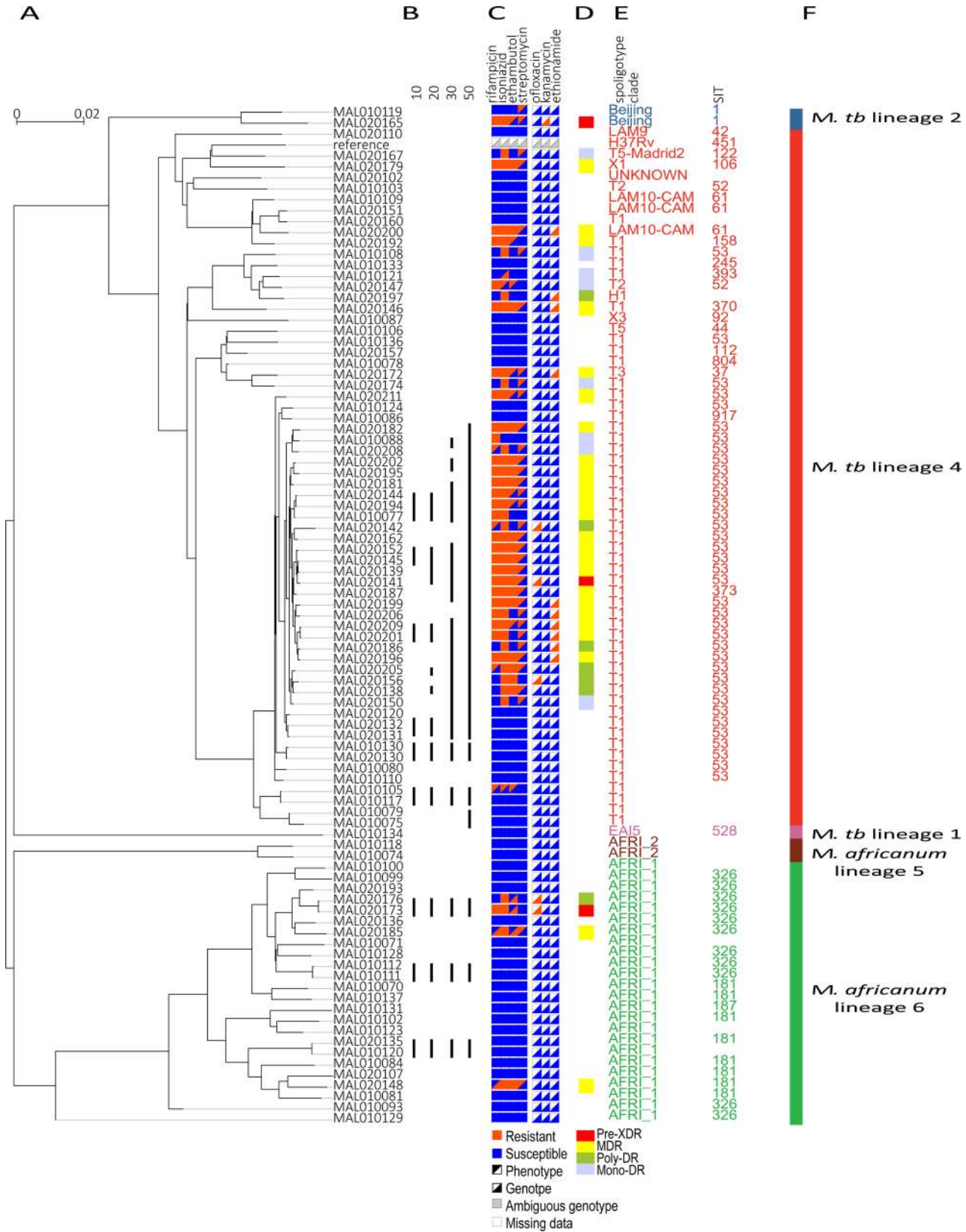
Our collection of 92 clinical MTC strains was isolated from patients presenting with pulmonary tuberculosis at Point G, Bamako, Mali between 2006 and 2010 as part of a cross-sectional

study to analyze the diversity of the MTC in Mali [20]. All patients were Mali natives, with the exception of one patient born in central Africa (S1A Table). We sequenced this collection using the Illumina sequencing platform, and the resulting reads were both assembled into contigs and aligned against the *M. tuberculosis* H37Rv reference genome. Based on our phylogenetic reconstructions, our collection included one lineage 1, two lineage 2, zero lineage 3, 63 lineage 4, two lineage 5 and twenty-four lineage 6 strains (Fig 1). The spoligotype distribution of our collection is representative of what has previously been observed in West Africa, except that we had a higher proportion of SIT53 (T1) strains and a lower proportion of SIT181 (AFRI\_1) strains (Figs 1 and S1) [45]. In order to perform statistical comparisons of the *M. tuberculosis*, *M. africanum* and *M. bovis* lineages, our newly sequenced dataset (the “Mali Collection”) was combined with data from additional strains from GenBank and South Africa (“Assembly Collection”, Fig 2), as well as data from China (“Alignment Collection”; see [Materials and Methods](#) and [S1 Table](#)). These additional comparator genomes enabled us to examine in detail the distinguishing characteristics of lineages 5 and 6 that might explain their geographic restriction.

Since this represents the largest collection of whole genome sequences of lineage 5 and 6 strains to date, we used our Assembly Collection to conduct a detailed examination of their phylogeny and characteristics in relation to other members of the MTC, including *M. bovis* and *M. tuberculosis*. *M. bovis* is considered an animal strain that mainly infects cattle and rarely humans, while *M. tuberculosis* is human adapted, and lineages 5 and 6 are thought to be intermediate between the two [1, 15]. Using our Assembly Collection, we constructed a high-resolution phylogenetic tree using 3,343 single-copy core orthogroups (sets of orthologs) conserved across all 137 strains (Materials and Methods). This tree was rooted using the outgroup *M. canettii* and agreed with phylogenies observed in other studies, including the fact that each of the lineages was clearly separated from the other, with lineage 5 being more closely related to human-adapted strains and lineage 6 being more closely related to *M. bovis* (Fig 2) [13, 15].

It has been previously shown, using average nucleotide identity (ANI) analysis, that separate bacterial species share <65–90% of genes and have no more than 94–95% ANI among shared genes [50, 51]. Using gene content and nucleotide variation among shared genes, we examined the genetic distances between strains within the Assembly Collection to understand how mycobacterial species fit within this framework. In agreement with previous studies showing the close relationship between MTC subspecies, including *M. africanum*, we observed that there was little diversity between the lineages analyzed [61]. Strikingly, values from inter-lineage comparisons of *M. tuberculosis*, *M. bovis*, and *M. africanum* strains overlapped those from intra-lineage comparisons, showing very little separation, with >99% ANI and >94% fraction of shared genes (Fig 3). These results are in agreement with previous observations that these different organisms should not, in fact, be named different species [61].

In contrast, MTC pairwise comparisons with *M. canettii* revealed a clear separation between the two groups, suggesting that they occupy distinct niches (Fig 3). *M. canettii* is a smooth tubercle bacilli that causes human tuberculosis in East Africa and is considered an emerging pathogen in some parts of the world, but its natural host(s) and reservoirs remain unknown [62]. Thus, it might be argued, based on these data and the traditional cutoffs set by ANI analysis, that all MTC members should be named the same species, and that even *M. canettii* should be included since pairwise identities with MTC exceeded these thresholds (Fig 3). However, as Smith et al. have previously discussed [61] changes in nomenclature can cause confusion in the literature, and so we will continue to refer to *M. africanum*-associated lineages as either lineage 5 or 6 within the MTC.





**Fig 1. *M. africanum* and *M. tuberculosis* drug resistance is genetically similar.** A) SNP-based phylogenetic tree of 92 newly sequenced strains from Mali (the Mali Collection), constructed using FastTree [55]. B) Groups differing by 10, 20, 30, or 50 SNPs are connected with black bars, as calculated in Cohen et al. [25]. C) Comparison of genotypic and phenotypic DST (drug susceptibility testing). Genotypic drug resistance was calculated using a list of mutations known to confer drug resistance (S2A Table) [59, 60]. D) Drug resistance category (mono-DR, poly-DR, MDR, or pre-XDR) based on genotype. E) Digital spoligotype clade and F) digital spoligotype international type (SIT), colored by lineage. F) The SIT is blank if the SITVITWEB database [45] did not contain a SIT for that strain's digital spoligotype pattern.

doi:10.1371/journal.pntd.0004332.g001

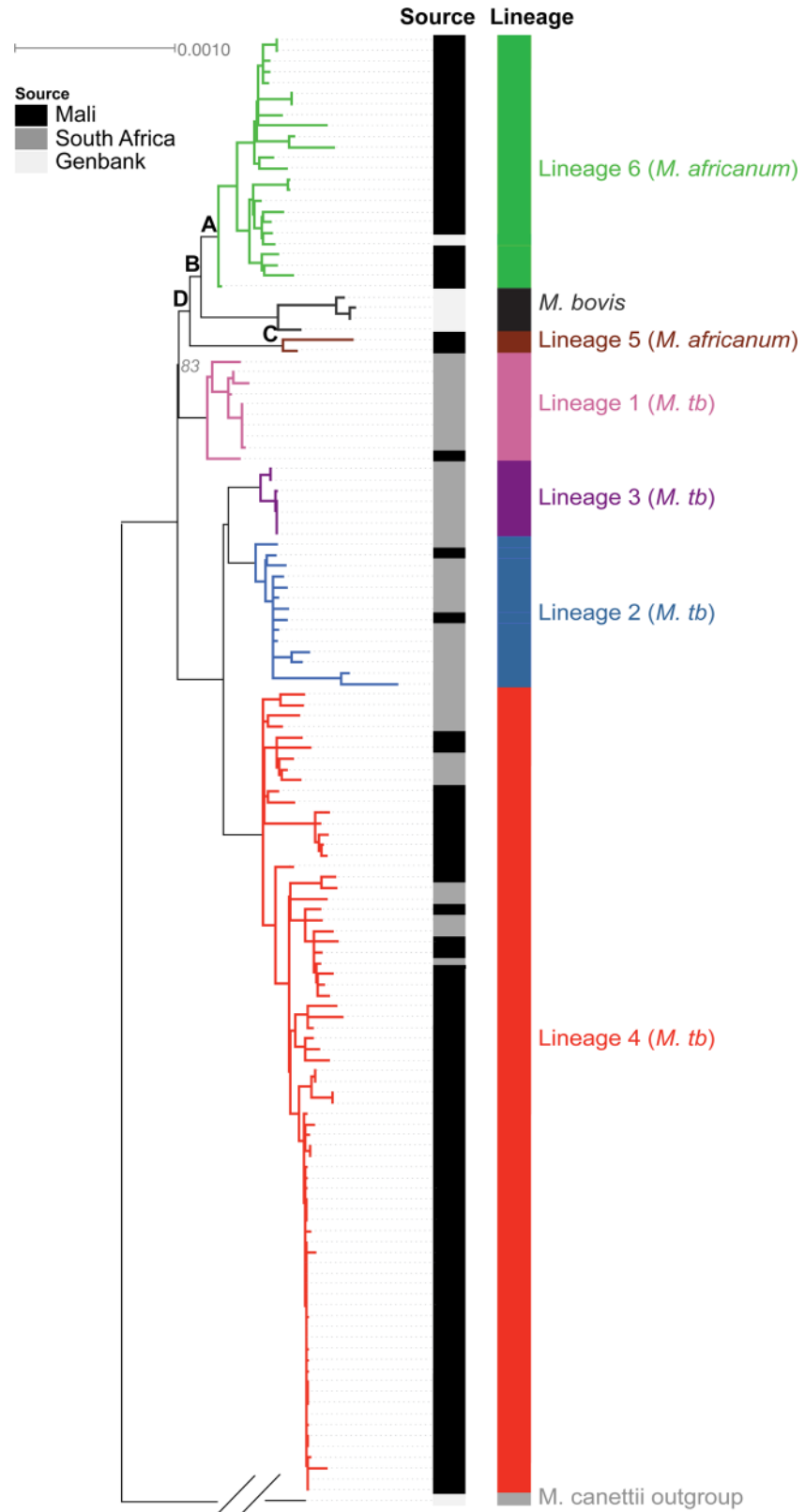
## Lineage 6 is as diverse as lineage 4 strains and involved in recent person-to-person transmission

Despite the fact that lineages 5 and 6 are so closely related to lineages 1–4, as demonstrated by Fig 3, they are still unique in being geographically restricted compared to these other lineages. One hypothesis for this restriction is that they are less fit, unable to compete with other lineages within the MTC. To examine this possibility, we looked within the Mali Collection for clues that lineage 6 strains were not as diverse as strains from lineage 4, the other predominant lineage within the region. We analyzed the breadth of pairwise diversity within lineage 6 using the ANI output and compared this diversity to that of lineage 4 strains isolated within Mali. ANI diversity was not statistically different when comparing these two groups of strains (Fig 4). Although this result does not eliminate the possibility of differing ecologies, such as an animal reservoir for lineage 6, as has previously been hypothesized [16], it does suggest that lineage 6 has not undergone a recent selective sweep or population bottleneck that would make lineage 6 populations circulating within Mali less diverse than lineage 4 populations [63].

In addition to being diverse, we also observed highly similar lineage 6 strains among this collection. Three pairs of lineage 6 strains were separated by less than 10 SNPs relative to *M. tuberculosis* H37Rv (Fig 1B; see Materials and Methods), including isolates from both HIV-positive and immunocompetent patients. There were also six such clusters within lineage 4. A cutoff of 12 SNPs has previously been used to determine recent transmission [64]. Thus, strains separated by less than 10 SNPs provide evidence of transmission, suggesting that 6 of 24 (25%) of our lineage 6 strains and 13 of 63 (21%) of our lineage 4 strains were involved in recent transmission events, confirming previous observations based on alternative genotyping approaches that there is robust ongoing transmission of lineage 6 within this region [9].

## Lineages 5 and 6 are not enriched for mutations in genes associated with virulence

Given the reports of lineages 5 and 6 strains having decreased virulence [7, 9–11], we hypothesized that altered virulence may contribute to geographical restriction, either due to changes in host requirements or to a reduction in fitness. To test this hypothesis, we examined lineage-specific pseudogenes (truncated genes) and non-synonymous SNPs in known essential genes, slow growth genes, and genes required for virulence in mice and growth in macrophages to determine whether lineages 5 and 6 had an enrichment of defects in these genes that might contribute to overall altered virulence [65–67]. Although both lineages 5 and 6 had lineage-specific mutations in these gene categories, so did other lineages (S4A and S5 Tables), and the proportion of mutated genes in lineage 6 was not significantly different from that of the other MTC lineages [8] (Fig 5). Lineage 4 was not included on this graph because it only had one lineage-specific mutation in an intergenic region when aligned to *M. tuberculosis* H37Rv, which is a member of lineage 4, and lineage 5 was excluded due to low sample size. We performed a similar analysis on the full length of genes encoding known T cell antigens as defined by Comas et al. [4] to explore whether alterations in these genes might be restricting host specificity, but again we observed no significant difference in the proportion of lineage 6-specific



**Fig 2. Phylogenetic tree of 92 newly sequenced strains from Mali, together with 45 additional strains with whole-genome assemblies (the Assembly Collection). Nodes, lineages, and newly-sequenced Mali**

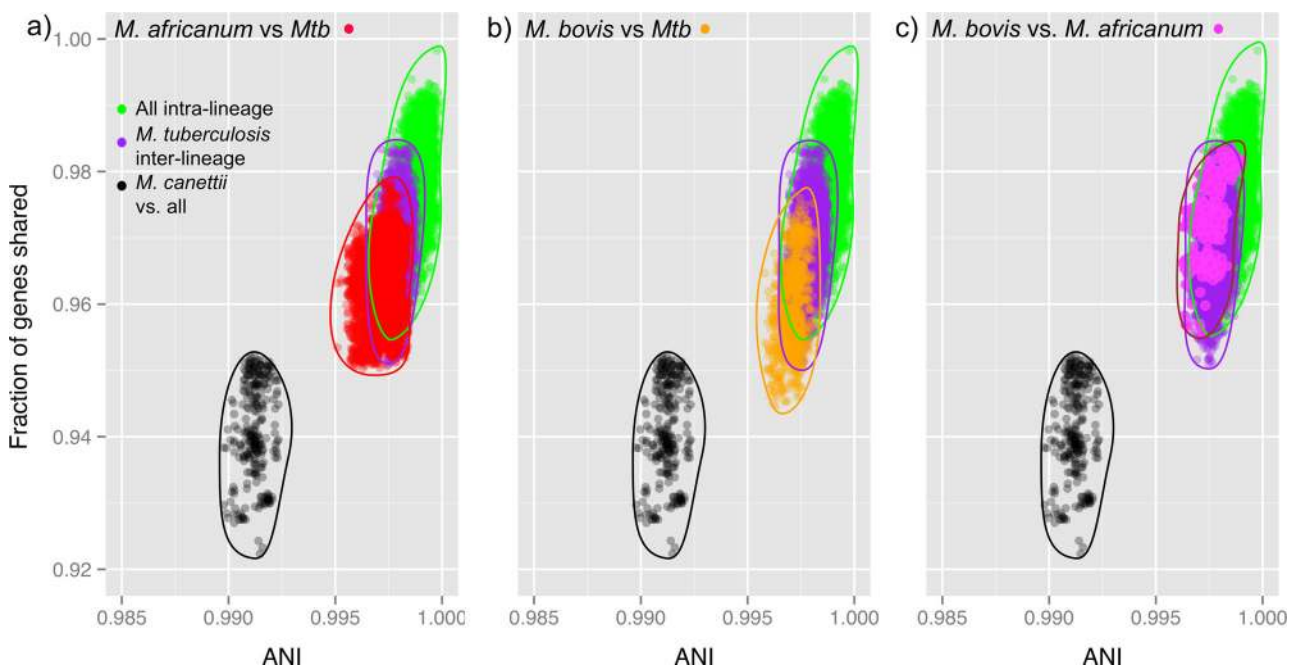
strains are indicated. All key nodes separating the major lineages had bootstrap values of 100%, except for the node separating *M. tuberculosis* lineage 1 and *M. africanum* lineage 5, which had a bootstrap value of 83%. Letters indicate key nodes analyzed in detail: (A) lineage 6, (B) the clade including *M. bovis* and lineage 6, (C) lineage 5, and (D) the clade including lineages 5, 6 and *M. bovis*.

doi:10.1371/journal.pntd.0004332.g002

mutations that fell within these genes as compared to lineages 1, 2 and 3 (Fig 5). Similarly, we looked for enrichment of lineage-specific mutations in COG, GO, KEGG, Pfam and TIGRFam gene categories, but found no enrichment in any of these categories, either for pseudogenes or non-synonymous SNPs ( $Q > 0.05$ ). These results corroborate our observations from ANI that the lineages of the MTC are very similar in their overall genetic composition and suggest that lineage 6 is not enriched for mutations in virulence genes relative to other lineages. However, while the overall number of mutations in virulence genes was not enriched, we identified mutations in these genes that might have an impact on virulence that will be discussed below.

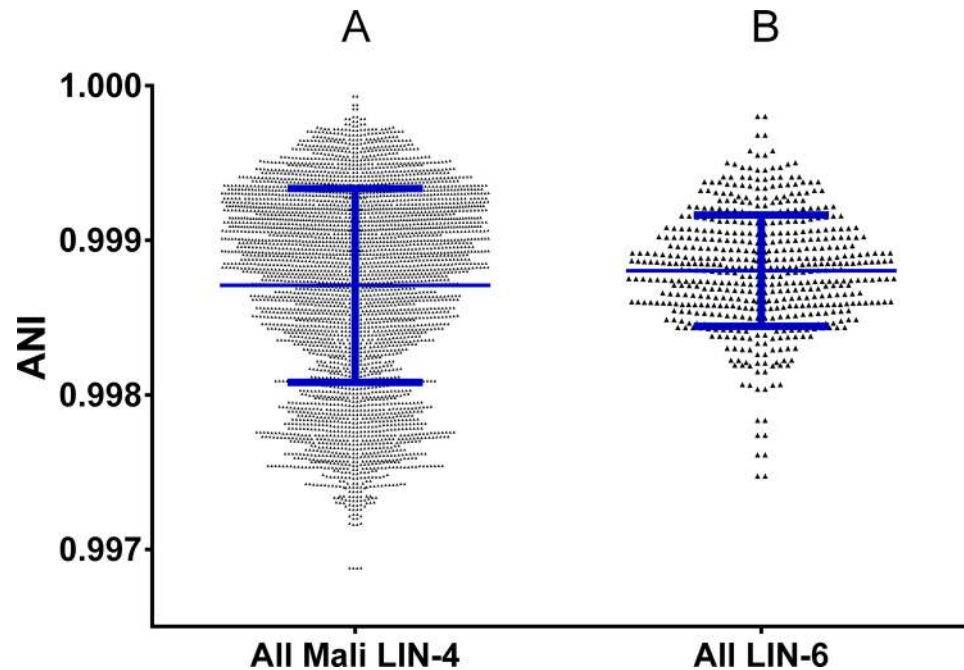
### Lineage 6 evolves drug resistance through similar mechanisms to other MTC lineages

Studies have shown that lineages 5 and 6 evolve drug resistance less often compared to other MTC lineages, including the study from which these sequenced strains were obtained [20, 68]. Thus, one hypothesis for the limited geographic range of lineages 5 and 6 could be decreased fitness relative to strains better able to evolve antibiotic resistance. In this case, we might expect that mutations driving drug resistance in these two lineages would be different from those evolving in more successful lineages. Thus, we analyzed our newly sequenced strains from Mali



**Fig 3. Average nucleotide identity (ANI) analysis indicates *M. africanum* and *M. tuberculosis* are not separate species.** A) ANI values when comparing *M. africanum* and *M. tuberculosis* do not cross the ANI species threshold of 94–95%. In fact, this comparison shows that the distribution of *M. africanum*/*M. tuberculosis* comparisons (red) overlaps that of inter-lineage *M. tuberculosis* comparisons (purple), indicating that *M. africanum* should be considered another lineage of *M. tuberculosis*. B) Similarly, ANI values when comparing *M. bovis* and *M. tuberculosis* also overlap with inter-lineage *M. tuberculosis*, and indicate that *M. bovis* should also be considered another lineage of *M. tuberculosis*. C) ANI values comparing *M. africanum* and *M. bovis* (pink) also overlap inter-lineage *M. tuberculosis* comparisons (green).

doi:10.1371/journal.pntd.0004332.g003



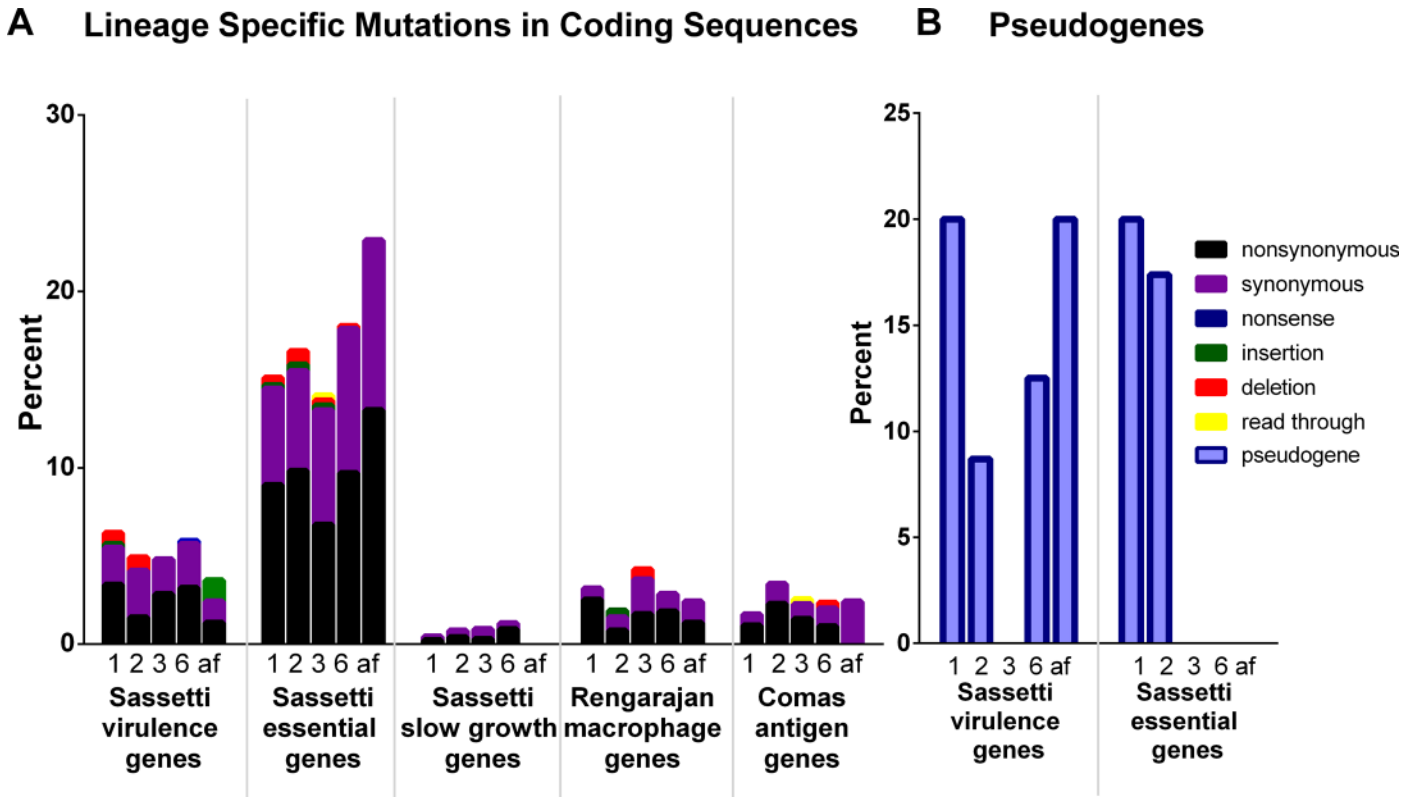
**Fig 4. Diversity in Mali lineage 4 and lineage 6 strains.** ANI values for comparisons (A) within all Mali lineage 4 isolates and (B) within all lineage 6 isolates. Blue lines indicate mean  $\pm$  standard deviation. The means of these two groups was not significantly different using the Mann-Whitney test.

doi:10.1371/journal.pntd.0004332.g004

for the presence of mutations known to confer drug resistance and used in common nucleic acid-based commercial tests [59, 60] for the detection of drug resistance [69–75] (S2A Table). Forty (60%) strains in lineages 1–4 and only four (15%) of the lineage 5 and 6 strains were phenotypically resistant to at least one of the four tested drugs. We observed that mutations used in commercial tests were sensitive in detecting phenotypic resistance to rifampicin, isoniazid and ethambutol (S2B Table and Figs 1 and S1).

Streptomycin resistance mutations were not included among our list of known resistance mutations (S2A Table). Therefore, we searched for potential resistance mutations in a set of genes previously known to affect streptomycin resistance, including *rrs*, *rpsL*, and *gidB* [76–78]. We identified a point mutation in *gidB* that caused a non-synonymous change (leucine to serine at residue 79) that is predicted to affect protein function [57] (S1D Fig; see S1 Text for more details). This mutation was found in 23 streptomycin resistant strains and no streptomycin susceptible strains in our dataset and likely represents a previously uncharacterized mutation that confers resistance to this drug. Previous studies have identified loss of function mutations in *gidB* affecting streptomycin resistance [77], as well as point mutations in the region of *gidB* close to residue 79, including at residues 75 and 82 [78].

In addition, we identified known mutations in genes associated with resistance to drugs that were not phenotypically assessed, including ofloxacin, kanamycin, and ethionamide. Using the list of mutations in S2A Table, we found that 25 (38%) of the Mali strains belonging to lineages 1, 2 or 4 could be classified as MDR (multi-drug resistant; resistant to isoniazid and rifampicin), and two (3%) could be classified as pre-XDR (pre-extensively drug resistant; resistant to isoniazid, rifampicin, plus either ofloxacin or kanamycin). In contrast, three (11%) of the lineage 5 and 6 strains could be classified as MDR, and one (4%) could be classified as pre-XDR. The presence of these pre-XDR strains is of particular concern, as XDR has not been reported in Mali, and testing is not currently performed routinely for second line antibiotics [79, 80].



**Fig 5. Percentage of lineage-specific mutations in virulence associated genes.** A) Percentage of lineage-specific mutations in coding sequences of the genes in each category. Sassetti virulence genes are genes that were identified in [66] as being required for virulence in mice. Sassetti essential and slow growth genes were identified by Sassetti et al. under *in vitro* conditions using TraSH [65]. Rengarajan macrophage genes were identified by Rengarajan et al. as being required for growth in macrophages [67]. Comas antigen genes were genes identified by Comas et al. as containing T cell epitopes [4]. The color of the bar indicates type of mutation. B) Percentage of lineage-specific pseudogenes falling into the above defined categories. Missing categories had no pseudogenes in any lineage. Lineage is indicated by the number below each bar, while 'af' indicates mutations found in both lineages 5 and 6 (both *M. africanum* lineages).

doi:10.1371/journal.pntd.0004332.g005

Similar resistance-conferring mutations were found among the lineages (S2 Fig). Although we cannot eliminate the possibility of cross resistance and other alternate genetic mechanisms of lineage 5 and 6 drug resistance, or of differences in drug tolerance or rates of persister cells, it appears that the mechanism of genetic drug resistance was similar between lineages 2, 4, 5 and 6. Thus, although the sample size was small, our results suggest that drug resistance, while less frequent in lineage 6, evolves through acquisition of similar mutations as observed in lineages 2 and 4 in Mali, including combinations of mutations leading to pre-XDR, and that this resistance could be detected using current molecular diagnostic approaches.

### Individual lineage-specific features suggest additional mechanisms that could be involved in geographic restriction

Previous analyses pinpointing lineage 6-specific genomic features have compared limited numbers of strains, which might have caused these studies to miss important features or to identify features that are not actually found in a broader set of lineage 6 strains [8, 19]. Also, these studies have not examined genomes of lineage 5 in detail. Using both our Alignment and Assembly Collections, containing representatives from lineages 1 through 6 and *M. bovis*, we sought to robustly identify distinguishing features of lineages 5 and 6, focusing on traits that could have caused geographic restriction. Using our Assembly Collection, at each node labeled A-D in Fig

2 (representing genetic diversification events that may correlate with ecological specialization), we identified gene gains and losses (Table 1; Materials and Methods). Many of our findings agreed with previous observations describing regions of difference, as determined through genomic hybridizations [3, 18]. However, we also identified a small number of genes that were not previously identified as being part of these known regions of difference (see S1 Text for full details), including a gain of genes encoding a PE-PGRS and hypothetical protein at the last common ancestor of lineage 6 and *M. bovis*, the loss of *Rv1523* (a methyltransferase) and *Rv3514* (*PE-PGRS57*) in lineage 5, and the loss of a gene encoding a TetR family regulator and the gain of one PPE protein-encoding gene at the last common ancestor of lineages 5, 6 and *M. bovis*.

In addition, using our alignments to *M. tuberculosis* H37Rv, we identified a number of lineage-specific mutations, including pseudogenes that affect protein function (Tables 2, S4, S5 and S6). From these data, we identified 681 lineage 6-specific mutations shared across all lineage 6 strains, including eight truncated pseudogenes. These data also provided the first in-depth analysis of lineage 5 assemblies, which revealed 952 lineage-specific mutations and 43 pseudogenes as shared by our two lineage 5 strains (see S1 Text). The larger number for lineage 5 compared to other lineages likely results from our small sample size. Key categories of lineage-specific mutations and pseudogenes that might contribute to the geographic restriction of lineages 5 and 6 are discussed below, and in more detail in the S1 Text.

### Mutations affecting ESX secretion could contribute to differing immune responses elicited by lineages 5 and 6

One distinguishing clinical characteristic of lineage 6 is an attenuated T cell response to ESAT-6, one of the proteins secreted through the ESX secretion system, as compared to patients infected with lineages 1–4 [5]. This altered immune response supports the hypothesis that lineage 5 and 6 have specificity for a particular host immunogenic background. Although our data cannot address whether ESAT-6 production has been affected, we observed non-synonymous polymorphisms, including indels, in genes encoding ESX secretion systems that could contribute to the different immune responses of lineage 6-infected patients (Table 3). Furthermore, we observed lineage-specific mutations in ESX-encoding genes in all lineages, suggesting that each lineage may have unique interactions with the host (Table 3; S1 Text).

### Alterations in cofactor biosynthesis pathways could impact many cellular functions in lineages 5 and 6

Lineages 5 and 6 had lineage-specific mutations, including pseudogenes, in genes encoding multiple components of cofactor biosynthetic pathways, including molybdenum, vitamin B12, and vitamin B3 (S1 Text and Tables 3, S4 and S5). Molybdenum cofactors are key catalysts for redox reactions, and are hypothesized to have played an important role in the evolution of pathogenic mycobacteria [81]. In addition, mycobacteria are one of the few bacterial pathogens with the ability to synthesize vitamin B12 [82]. Thus, both of these cofactors have specifically evolved in mycobacteria and loss of these cofactor biosynthetic pathways could affect the function of proteins that use these cofactors, which include proteins that are important for many cellular functions. These mutations may affect the host range of lineages 5 and 6, supporting the hypothesis of a unique host preference.

**Table 1. Orthologs identified in gene content analysis as lost or gained at nodes A-D.** For some regions of difference (RD), the first and/or last gene in the region was not identified in our analysis because enough of the gene remained to align to H37Rv, and thus was not considered absent. Gene names starting with an O were identified through our SYNERGY orthogroup analysis. Nodes A-D were identified in Fig 2.

A: root node of lineage 6			C: root node of lineage 5		
loss	Annotation	RD	loss	Annotation	RD
Rv0124	PE-PGRS family protein PE_PGRS2	RD701	Rv1334	hydrolase	RD711
B: root node of lineage 6 and <i>M. bovis</i>			Rv1335	sulfur carrier protein CysO	RD711
loss	Annotation	RD	Rv1523	methyltransferase	
Rv0222	enoyl-CoA hydratase EchA1	RD10	Rv1978	hypothetical protein	RD713
Rv1965	ABC transporter permease YrbE3B	RD7	Rv1979c	permease	RD713
Rv1966	MCE-family protein Mce3A	RD7	Rv1993c	hypothetical protein	RD743
Rv1967	MCE-family protein Mce3B	RD7	Rv1994c	ArsR family transcriptional regulator CmtR	RD743
Rv1968	MCE-family protein Mce3C	RD7	Rv1995	hypothetical protein	RD743
Rv1969	MCE-family protein Mce3D	RD7	Rv3514	PE-PGRS family protein PE_PGRS57	
Rv1970	MCE-family lipoprotein LprM	RD7	D: root node of lineages 5 and 6 and <i>M. bovis</i>		
Rv1971	MCE-family protein Mce3F	RD7	loss	Annotation	RD
Rv1972	MCE-associated membrane protein	RD7	Rv2073c	oxidoreductase	RD9
Rv1973	MCE-associated membrane protein	RD7	Rv2074	pyridoxamine 5'-phosphate oxidase	RD9
Rv1974	membrane protein	RD7	Rv2084	hypothetical protein	
Rv1975	hypothetical protein	RD7	gained	Annotation	RD
Rv1976c	hypothetical protein	RD7	O850451604	PPE family protein	
Rv3617	epoxide hydrolase EphA	RD8			
Rv3618	monooxygenase	RD8			
Rv3619c	ESAT-6 like protein EsxV	RD8			
Rv3620c	ESAT-6 like protein EsxW	RD8			
Rv3621c	PPE family protein PPE65	RD8			
Rv3622c	PE family protein PE32	RD8			
gained	Annotation	RD			
O850447630	hypothetical protein				
O850450572	PE-PGRS family protein				

doi:10.1371/journal.pntd.0004332.t001

**Table 2. Summary of the lineage-specific mutations and pseudogenes detected for each lineage.**

Lineage	Mutations		Pseudogenes	
	Average ± SD	Lineage Specific	Average ± SD	Lineage Specific
LIN-1	2527.2 ± 39.2	536	189 ± 21.5	5
LIN-2	1774.4 ± 64.9	308	183.5 ± 43.1	23
LIN-3	1740.0 ± 52.9	406	157.9 ± 46.5	5
LIN-4	1050.0 ± 143.9	1	99.5 ± 28.8	0
LIN-5	2540.5 ± 4.9	952*	190.5 ± 0.7	43*
LIN-6	2605.8 ± 82.0	681	201.5 ± 26.4	8
LIN-5 and LIN-6	2600.4 ± 80.5	90	200.5 ± 25.4	5
LIN-1, LIN2, LIN-3, and LIN-4	1355.4 ± 400.0	0	148.4 ± 55.4	0
LIN-1, LIN2, and LIN-3	1803.4 ± 161.9	NA	182.3 ± 42.5	NA

\*Due to low lineage 5 sample size (only two lineage 5 strains), there is low confidence in which mutations were lineage 5-specific.

doi:10.1371/journal.pntd.0004332.t002

**Table 3. Summary of lineage-specific mutations highlighted in the results and discussion.** Each cell indicates the number of lineage-specific mutations in that category (see [Materials and Methods](#)). “Indel” indicates an insertion or deletion; “synonymous” indicates a SNP that does not change the amino acid sequence; “non-synonymous” indicates a SNP that alters the amino acid sequence; and “non-synonymous and affects protein function” is the subset of non-synonymous SNPs that were predicted by SIFT [57] to affect protein function. Each column heading refers to the gene category, each of which is discussed in more detail in the results section. Due to low lineage 5 sample size (only two lineage 5 strains), there is low confidence in which mutations were lineage 5-specific.

Category:	ESX Secretion	molybdenum, riboflavin and cobalamin	L,D transpeptidase	adenylate cyclase	drug-resistance associated genes	mammalian cell entry (MCE)
Lineage 1	Indel	0	1	0	0	0
	Synonymous	3	2	0	1	0
	Non-synonymous	4	3	0	1	1
	Non-synonymous & affects protein function	1	2	0	0	1
Lineage 2	Indel	1	1	0	0	0
	Synonymous	2	2	0	0	0
	Non-synonymous	7	0	0	0	0
	Non-synonymous & affects protein function	3	0	0	0	0
Lineage 3	Indel	0	0	0	0	0
	Synonymous	2	1	0	2	1
	Non-synonymous	3	0	0	0	1
	Non-synonymous & affects protein function	0	0	0	0	1
Lineage 5	Indel	3	1	0	0	1
	Synonymous	10	7	0	2	1
	Non-synonymous	16	8	1	2	2
	Non-synonymous & affects protein function	7	4	1	2	2
Lineage 6	Indel	2	0	0	1	0*
	Synonymous	1	7	0	0	0
	Non-synonymous	7	7	2	2	3
	Non-synonymous & affects protein function	1	3	1	0	0
Lineage 5 & 6	Indel	0	2	0	0	0
	Synonymous	2	0	0	0	0
	Non-synonymous	0	0	0	0	0
	Non-synonymous & affects protein function	0	0	0	0	0

\*Table does not include the 8 genes completely deleted in RD7

doi:10.1371/journal.pntd.0004332.t003

### Mutations that could cause distinct cell size, morphology and growth of lineages 5 and 6

It has been shown previously that *M. africanum* GM041182 has a distinct physiology as compared to that of *M. tuberculosis* H37Rv, including a larger cell size and slower growth rate [7]. Possibly explaining these differences, we identified lineage 6-specific non-synonymous SNPs in



genes encoding the L,D transpeptidases, *ldtA* and *ldtB* (*Rv0166c* and *Rv2518c*), previously shown to form cross-linkages within peptidoglycan (Tables 3 and S4A) [83] and to be key drivers of cell shape, size, surface morphology, growth and virulence [84]. Lineage 5 also contained a non-synonymous SNP predicted to affect LdtA protein function (Tables 3 and S4A). No other lineages had a lineage-specific mutation in an L,D-transpeptidase.

### Mutations affecting cAMP signaling in lineages 5 and 6 could affect virulence

We observed that lineages 5 and 6 had lineage-specific mutations in genes encoding adenylate cyclases, the enzymes that synthesize cyclic AMP (cAMP), an important cell signaling molecule. Although the affected genes were different between the two lineages, no other lineage had lineage-specific mutations predicted to affect adenylate cyclase function. Deletion of one of the 17 adenylate cyclases in *M. tuberculosis*, *Rv0386*, has been shown to reduce virulence and alter the immune response [85]. Bentley et al. also previously found that this gene was a pseudogene in *M. africanum* GM041182, although here we find that pseudogenization of *Rv0386* was not lineage specific (S6A Table). Nevertheless, given the number of affected adenylate cyclases, there may be differences in cAMP signaling within lineages 5 and 6, leading to altered pathogenicity.

### Distinct mutations in drug resistance-associated genes in lineages 5 and 6

In order to shed light on the reported lower rates of drug resistance in lineages 5 and 6, we screened our lineage-specific mutations to investigate if there were any changes in known drug resistance genes that were not on the list of mutations used before and that might affect the development of antibiotic resistance [20, 68]. In lineage 6, we observed two lineage-specific non-synonymous mutations in *rpoB*, and one lineage-specific non-synonymous mutation in *embC* (S1A and S1C Fig and Tables 3 and S4A) not previously implicated in antibiotic resistance. Lineage 5 strains had non-synonymous mutations in genes encoding AtpH (*Rv1307*) and AtpG (*Rv1309*), both of which are subunits of ATP synthase [86] (S4A Table), and a target of bedaquiline, a new antibiotic reserved for the treatment of drug resistant tuberculosis [87]. Both of these mutations were predicted to affect protein function by SIFT [57], and may affect bedaquiline efficacy in countries with a high proportion of patients infected with lineage 5. Thus, both of these lineages have non-resistance conferring mutations in genes associated with drug resistance that might influence the frequency at which drug resistance develops in these lineages.

### Mutations in mammalian cell entry genes that could affect virulence in lineages 5 and 6

*M. tuberculosis* H37Rv contains four mammalian cell entry (MCE) operons, which play an important role in mycobacterial virulence [88]. In addition to confirming earlier reports that lineage 6 strains lacked one of these four operons (operon 3; Table 1) [18], we observed lineage-specific mutations in several of the other MCE operons (lineage 6 had mutations in operons 1 and 2; lineage 5 had mutations in operons 1 and 3). We also observed a non-synonymous mutation in *mce1B* that was shared by lineages 5 and 6 strains and was predicted by SIFT to affect Mce1B protein function [57]. In comparison, the other lineages had nearly identical MCE operons as compared to *M. tuberculosis* H37Rv (Tables 3 and S4A).

## Discussion

Our study describes the largest collection of sequenced lineage 6 isolates to date, and, to our knowledge, the first in-depth analysis of the genetics of lineage 5. Through our work, we have characterized the genetic basis of antibiotic resistance in lineage 6 strains from Mali, shown that *M. africanum* and *M. tuberculosis* are part of the same species, and better defined the mutations and changes in gene content that typify these lineages. Collectively, this work provides insights into these understudied lineages and provides testable hypotheses as to why they are geographically restricted.

We evaluated 92 Mali MTC isolates using both assembly and alignment-based approaches. Our assemblies revealed several new regions of difference and our alignments identified smaller lineage-specific changes. In addition to our conclusion that *M. africanum* is not a separate species, we observed that some *M. africanum*-*M. tuberculosis* pairs of strains have greater average nucleotide identity than some pairs of strains from within the same lineage. Furthermore, our ANI data demonstrated that there is comparable diversity in lineages 4 and 6, suggesting that lineage 6 has not undergone a recent population bottleneck. This emphasizes the extremely close relationship between all MTC lineages, highlighting the role that small changes within the MTC have played in geographical restriction and altering host preferences. Since our assemblies were of very high quality, we were able to observe changes in genes that previous studies could not, thus providing a prioritized list of genes for investigating lineage 5 and 6 characteristics.

One hypothesis for the geographical restriction of lineages 5 and 6 is the presence of an unknown non-human reservoir. *M. africanum* has been found in animals, including monkeys, cows, pigs and hyrax [89–94]. Unfortunately, given genomic data from human clinical isolates alone, we cannot address this hypothesis directly. However, given the similar level of diversity between lineage 4 and 6 in Mali and the evidence of person-to-person transmission, a non-human reservoir seems unlikely to explain the geographic restriction, as lineage 6 appears well adapted to spread in humans living in this geographic setting, unlike *M. bovis* in this and other settings [95–97].

Another hypothesis for why lineages 5 and 6 occur almost exclusively in West Africa is a preference for hosts of West African ethnicity, supported by previous evidence, including a study linking lineage 5 to the Ewe ethnic group [17]. We identified lineage-specific mutations in ESX genes in every lineage, indicating that each lineage may interact uniquely with the host immune system. Mycobacteria have five ESX secretion systems, also known as type VII secretion systems, which secrete small proteins across the bacterial cell envelope and are important to mycobacterial virulence [98, 99]. For example, ESX-1 secretion is lost as part of RD1 in *M. bovis* BCG vaccine strains, resulting in loss of ESAT-6 and CFP-10 secretion, and thus attenuation of the bacterium [100, 101]. The lineage-specific mutations in ESX genes could lead to alterations in the pathogen-host immune interaction, resulting in a requirement in lineages 5 and 6 for the West African immune system. In fact, an altered response to ESAT-6 in patients infected by lineage 6 has previously been reported [5]. Thus, the specific ESX mutations in lineages 5 and 6 could represent adaptations to the niche of the West African host.

Lineage-specific mutations in cobalamin biosynthesis could also contribute to adaptation of these lineages to the specific ecological niche of the West African host. The hypothesis of adaptation to a different host cofactor environment for lineages 5 and 6 is supported by several studies that have found increased levels of vitamin B12 plasma concentrations in West Africans compared to Europeans and Mexicans [102, 103]. One unique characteristic of mycobacteria compared to many other bacterial genera is that they are capable of synthesizing vitamin B12. Furthermore, vitamin B12 may play a crucial role in *M. tuberculosis* infection [82, 104, 105].

Thus, the lineage-specific mutations in cobalamin pathways in lineages 5 and 6 may alter these strains' ability to synthesize vitamin B12, which may be tolerated in West African hosts with higher levels of plasma B12. Adaptation to this B12-rich West African niche might prevent these lineages from infecting other ethnic groups with lower B12 bioavailability; however, further studies would be required to confirm this hypothesis.

A third hypothesis for the geographic restriction of lineages 5 and 6 is that they are less fit, either for transmission or in-host virulence, resulting in a decreased ability to survive outside of West Africa. Several papers have shown no difference in transmission rates between *M. tuberculosis*-associated strains and *M. africanum*-associated strains [5, 9, 106, 107]. In agreement with previous findings, our Mali Collection revealed three pairs of lineage 6 strains separated by 10 or fewer SNPs when aligned to *M. tuberculosis* H37Rv, suggesting recent transmission of strains between patients within the ethnic backgrounds prevalent in Mali [64]. That these transmission events were not exclusive to HIV positive patients suggests that a compromised immune system is not required for a transmission event. These results indicate that lineages 5 and 6 do not have a reduced ability to transmit.

A decrease in fitness could also be reflected in a decrease in virulence. It has been hypothesized that *M. africanum* is less virulent within humans, mice and guinea pigs than is *M. tuberculosis* [7, 9–11]. However, lineage 6 was not enriched for mutations in virulence and growth-related genes compared to lineages 1, 2 and 3, suggesting that lineage 6 does not contain an overall numerical loss of virulence or growth-associated genes. Despite this, individual mutations can still greatly affect disease outcome, and analysis of lineage-specific mutations identified several potential mechanisms that could lead to changes in how lineages 5 and 6 proliferate and cause disease. The lineage-specific mutations discussed above that could relate to a niche adaptation in hosts of West African ethnicity, including the lineage-specific mutations in ESX genes and cofactor biosynthesis genes, are also involved in virulence.

Another key set of virulence genes with lineage 5 and 6-specific mutations are the MCE operons. The MCE operons play an important role in the virulence of *M. tuberculosis*, particularly in mycobacterial growth in macrophages [67]. Antibodies to MCE1 proteins have been identified in patients [108], and operons 1–3 are required for virulence in mice [88]. Despite this apparent role in virulence, lineage 6 contains mutations that affect protein function in operons 1–3, while lineages 1–3 have nearly identical MCE operons to *M. tuberculosis* H37Rv, suggesting one potential mechanism of decreased virulence.

Another set of virulence-related genes with lineage-specific mutations are adenylate cyclases, which synthesize cAMP, an important second messenger [109]. *M. tuberculosis* encodes 17 adenylate cyclases, and deletion of one of them (*Rv0386*) has been shown to affect virulence and host response [85], highlighting the importance of this set of genes to pathogenicity. Both lineages 5 and 6 contained lineage-specific mutations predicted to affect the protein function of several adenylate cyclases, suggesting altered cAMP signaling in these strains, and a potential effect on the virulence of lineages 5 and 6.

Another pathway that affects bacterial growth and host response is the synthesis of the cell wall. Both lineage 5 and 6 contained lineage-specific mutations in L,D-transpeptidases. L,D-transpeptidases are critical to the structure of mycobacterial peptidoglycan and are involved in bacterial structure and growth [84], providing a possible explanation for the reported changes in cell size and doubling time in *M. africanum* GM041182 compared to *M. tuberculosis* H37Rv [7]. An altered cell wall could support either the hypothesis of decreased virulence, or suggest the need for a specific host immune system.

In addition, we saw high variability in PE, PPE and PE-PGRS genes, including changes in gene content. These repetitive regions are difficult to sequence and are often ignored, but may play a crucial role in antigenicity and the host-pathogen interaction [110, 111]. Using our high

quality assemblies and alignments, we were able to identify lineage-specific mutations in these genes, as well as altered gene content. These mutations highlight the possibility of a critical role for these proteins in host-pathogen interactions and emphasize the need for a more detailed analysis of these regions. Furthermore, there were also a number of mutated hypothetical proteins and proteins of unknown function, all of which may play a critical as yet undiscovered role.

In addition to exploring mechanisms of geographic restriction, we also identified mutations that may have clinical implications for the region. We found that in Mali, *M. africanum*-associated and *M. tuberculosis*-associated strains evolved antibiotic resistance through similar mutations, and thus standard line-probe assays can still be utilized in West Africa. However, we also found a *gidB* polymorphism not previously described which might account for much of the streptomycin resistance in Mali. Also of concern, we identified several cases of pre-XDR in Mali, suggesting that Mali may need to begin testing for XDR cases. Furthermore, we identified lineage 5 or 6-specific mutations that may affect the evolution of drug resistance, particularly bedaquiline. Thus, whole genome sequencing surveys like this one are useful in revealing new mechanisms for drug resistance, informing development of molecular diagnostics.

One weakness of our study was that we were limited in our sample size for lineage 5 and *M. bovis* strains. Our collection was not representative of *M. bovis* genomic diversity, as three of the four *M. bovis* strains in our analysis were attenuated *M. bovis* BCG vaccine strains. However, we only used the *M. bovis* strains in our ANI and gene content analysis, and required that any observations be consistent with wild-type *M. bovis* sequence, AF2122/97, and our results corroborated all previous findings of *M. bovis* regions of difference. Another weakness of our study was that our observations may be specific to Mali, since all lineage 5 and 6 isolates sequenced for our study were isolated in Mali, although these lineages are found throughout West Africa. However, our lineage 6 isolates were genetically diverse, and represented multiple spoligotypes, and our isolates from other lineages did not cluster separately on the phylogenetic tree from strains isolated from South African patients. Thus, our collection reflected substantial diversity and did not originate from a clonal outbreak. In fact, the study from which we selected our samples found a wide diversity of strains in Mali, which covered 55% of all known spoligotyped strains [20]. Furthermore, based on spoligotyping, many similar strains can be found in neighboring countries [54, 68, 112–115]. And, finally, studies that employ genomic data alone are insufficient to address causality. However, we believe that this in-depth genomics analysis of the neglected pathogen, “*M. africanum*”, provides a strong foundation from which causal relationships between lineage-specific variation and geographic restriction can be made.

This collection provides valuable insights into the distinguishing genomic features of *M. africanum*. Here, we have analyzed in detail the genomes of lineage 5 and 6 isolates from Mali and identified several potential genetic reasons for the geographical restriction of lineages 5 and 6, such as alterations in vitamin B12 pathways and genes associated with virulence, which provide a guide to future studies focusing on the effects of specific genes. Although we cannot specifically point to a single reason why these lineages are geographically restricted, we have found mutations that support the hypothesis of a unique requirement for a host of West African ethnicity and for the hypothesis of loss of bacterial fitness. These hypotheses are not mutually exclusive, and we anticipate that these observations will be able to inform and fast-track experiments on mycobacterial pathogenicity and virulence, particularly with regard to this unique member of the MTC.

## Supporting Information

**S1 Fig. Comparison of our Mali collection to the previously observed distribution of West African spoligotypes.** Percent of strains in our Mali collection (black bars) or in West Africa (gray bars) based on (A) the spoligotype international type (SIT) or (B) spoligotype clade. West African numbers are based on the SITVITWEB database [45]. SITs or clades present at less than 1% in either group were excluded from the figure.  
(PDF)

**S2 Fig. Mutations in drug-resistance associated genes.** Plots showing details of mutations identified in genes known to confer drug resistance. Light blue or red horizontal shaded bars indicate phenotypic sensitivity or phenotypic resistance, respectively, for the strain of interest. The corresponding vivid color in a particular box indicates the presence of the resistance mutation represented by that column. A) rifampicin B) isoniazid C) ethambutol D) streptomycin.  
(PDF)

**S1 Table. Samples used in our study.** A) List of Mali Collection samples, with patient information and digital spoligotype results. B) List of all 137 Assembly Collection strains, used for our assembly-based analyses, including 91 newly sequenced strains from Mali, 40 strains from South Africa [25], and six strains from Genbank. C) Sequence Read Archive identifiers for each of the 161 additional strains from China used in our SNP analysis as part of the Alignment Collection [27].  
(XLSX)

**S2 Table. Drug resistance analysis.** A) Drug resistance mutations analyzed. B) Table showing true positives, false positives, true negatives, false negatives, sensitivity, and specificity for the four drugs for which we have phenotype information for all strains.  
(XLSX)

**S3 Table. Schematic overview of key metrics for lineage specificity calculations.**  
(DOCX)

**S4 Table. Lineage-specific mutations.** All lineage-specific mutations (A) in coding sequences and (B) in intergenic regions. Maf indicates mutations shared between lineages 5–6, but not found in lineages 1–4. No mutations were shared between lineages 1–4 but not lineages 5–6.  
(XLSX)

**S5 Table. Lineage-specific pseudogenes.** Maf indicates mutations shared between lineage 5 and 6 but not found in lineages 1–4. No mutations were shared between lineages 1–4 but not lineages 5–6.  
(XLSX)

**S6 Table. Comparison of pseudogenes to previous analysis.** A) Comparison of pseudogenes identified differently by our study to those identified by Bentley et al [19]. This table compares lineage 4, lineage 6 or *M. africanum*-specific pseudogenes identified in our study to pseudogenes identified by Bentley et al. as belonging to lineage 4, lineage 6, lineage 6 and animal strains, or lineages 5–6 and animal strains. A “0” indicates that the gene is not a pseudogene in that strain, while “1” indicates that it is, and “2” indicates an ambiguous call. Genes with a light blue background were identified in this study and not by Bentley et al., while genes with a light green background were identified by Bentley et al., but not by this study, and genes with a purple background were identified by both studies. B) Table summarizing the differences between our study and Bentley et al. [19].  
(XLSX)

**S1 Text. Additional information on the novel streptomycin resistance mutation and on the lineage-specific mutations.**

(DOCX)

**Acknowledgments**

The authors would like to express their sincere gratitude to the patients in Mali, without which this study would not have been possible. The authors also gratefully acknowledge the collaborative efforts of the following individuals whose significant prior contributions allowed for this study: Lucia Alvarado, Sinéad B. Chapman, Susanna Hamilton, Clinton Howarth, Jeffrey D. Larimer, Matthew D. Pearson, Margaret E. Priest, Sarah K. Young, Bruce Walker, Jennifer Wortman and Qiandong Zeng. Additionally, the authors would like to acknowledge the Broad Sequencing Platform for assistance with data acquisition.

**Author Contributions**

Conceived and designed the experiments: KW MM WRB. Performed the experiments: KW MM BD BB MS SD. Analyzed the data: KW AMM TA TS CAD. Contributed reagents/materials/analysis tools: AME WRB. Wrote the paper: KW AMM AME.

**References**

1. de Jong BC, Antonio M, Gagneux S. *Mycobacterium africanum*—review of an important cause of human tuberculosis in West Africa. PLoS neglected tropical diseases. 2010; 4(9):e744. doi: [10.1371/journal.pntd.0000744](https://doi.org/10.1371/journal.pntd.0000744) PMID: [20927191](https://pubmed.ncbi.nlm.nih.gov/20927191/)
2. Castets M, Boisvert H, Grumbach F, Brunel M, Rist N. [Tuberculosis bacilli of the African type: preliminary note]. Revue de tuberculose et de pneumologie. 1968; 32(2):179–84. PMID: [4985104](https://pubmed.ncbi.nlm.nih.gov/4985104/)
3. Mostowy S, Onipede A, Gagneux S, Niemann S, Kremer K, Desmond EP, et al. Genomic analysis distinguishes *Mycobacterium africanum*. Journal of clinical microbiology. 2004; 42(8):3594–9. PMID: [15297503](https://pubmed.ncbi.nlm.nih.gov/15297503/)
4. Comas I, Chakravarti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nature genetics. 2010; 42(6):498–503. doi: [10.1038/ng.590](https://doi.org/10.1038/ng.590) PMID: [20495566](https://pubmed.ncbi.nlm.nih.gov/20495566/)
5. de Jong BC, Hill PC, Brookes RH, Gagneux S, Jeffries DJ, Otu JK, et al. *Mycobacterium africanum* elicits an attenuated T cell response to early secreted antigenic target, 6 kDa, in patients with tuberculosis and their household contacts. The Journal of infectious diseases. 2006; 193(9):1279–86. PMID: [16586366](https://pubmed.ncbi.nlm.nih.gov/16586366/)
6. Tientcheu LD, Sutherland JS, de Jong BC, Kampmann B, Jafali J, Adetifa IM, et al. Differences in T-cell responses between *Mycobacterium tuberculosis* and *Mycobacterium africanum*-infected patients. European journal of immunology. 2014; 44(5):1387–98. doi: [10.1002/eji.201343956](https://doi.org/10.1002/eji.201343956) PMID: [24481948](https://pubmed.ncbi.nlm.nih.gov/24481948/)
7. Bold TD, Davis DC, Penberthy KK, Cox LM, Ernst JD, de Jong BC. Impaired fitness of *Mycobacterium africanum* despite secretion of ESAT-6. The Journal of infectious diseases. 2012; 205(6):984–90. doi: [10.1093/infdis/jir883](https://doi.org/10.1093/infdis/jir883) PMID: [22301632](https://pubmed.ncbi.nlm.nih.gov/22301632/)
8. Gehre F, Otu J, DeRiemer K, de Sessions PF, Hibberd ML, Mulders W, et al. Deciphering the growth behaviour of *Mycobacterium africanum*. PLoS neglected tropical diseases. 2013; 7(5):e2220. doi: [10.1371/journal.pntd.0002220](https://doi.org/10.1371/journal.pntd.0002220) PMID: [23696911](https://pubmed.ncbi.nlm.nih.gov/23696911/)
9. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. The Journal of infectious diseases. 2008; 198(7):1037–43. doi: [10.1086/591504](https://doi.org/10.1086/591504) PMID: [18702608](https://pubmed.ncbi.nlm.nih.gov/18702608/)
10. de Jong BC, Adetifa I, Walther B, Hill PC, Antonio M, Ota M, et al. Differences between tuberculosis cases infected with *Mycobacterium africanum*, West African type 2, relative to Euro-American *Mycobacterium tuberculosis*: an update. FEMS immunology and medical microbiology. 2010; 58(1):102–5. doi: [10.1111/j.1574-695X.2009.00628.x](https://doi.org/10.1111/j.1574-695X.2009.00628.x) PMID: [20002176](https://pubmed.ncbi.nlm.nih.gov/20002176/)
11. Castets M, Sarrat H. [Experimental study of the virulence of *Mycobacterium africanum* (preliminary note)]. Bulletin de la Societe medicale d'Afrique noire de langue francaise. 1969; 14(4):693–6. PMID: [5407132](https://pubmed.ncbi.nlm.nih.gov/5407132/)

12. Meyer CG, Scarisbrick G, Niemann S, Browne EN, Chinbuah MA, Gyaopong J, et al. Pulmonary tuberculosis: virulence of *Mycobacterium africanum* and relevance in HIV co-infection. *Tuberculosis*. 2008; 88(5):482–9. doi: [10.1016/j.tube.2008.05.004](https://doi.org/10.1016/j.tube.2008.05.004) PMID: [18590979](https://pubmed.ncbi.nlm.nih.gov/18590979/)
13. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature genetics*. 2013; 45(10):1176–82. doi: [10.1038/ng.2744](https://doi.org/10.1038/ng.2744) PMID: [23995134](https://pubmed.ncbi.nlm.nih.gov/23995134/)
14. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, et al. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerging infectious diseases*. 2013; 19(3):460–3. doi: [10.3201/eid1903.120256](https://doi.org/10.3201/eid1903.120256) PMID: [23622814](https://pubmed.ncbi.nlm.nih.gov/23622814/)
15. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Seminars in immunology*. 2014; 26(6):431–44. doi: [10.1016/j.smim.2014.09.012](https://doi.org/10.1016/j.smim.2014.09.012) PMID: [25453224](https://pubmed.ncbi.nlm.nih.gov/25453224/)
16. Smith NH, Kremer K, Inwald J, Dale J, Driscoll JR, Gordon SV, et al. Ecotypes of the *Mycobacterium tuberculosis* complex. *Journal of theoretical biology*. 2006; 239(2):220–5. PMID: [16242724](https://pubmed.ncbi.nlm.nih.gov/16242724/)
17. Asante-Poku A, Yeboah-Manu D, Otchere ID, Aboagye SY, Stucki D, Hattendorf J, et al. *Mycobacterium africanum* Is Associated with Patient Ethnicity in Ghana. *PLoS neglected tropical diseases*. 2015; 9(1):e3370. doi: [10.1371/journal.pntd.0003370](https://doi.org/10.1371/journal.pntd.0003370) PMID: [25569290](https://pubmed.ncbi.nlm.nih.gov/25569290/)
18. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99(6):3684–9. PMID: [11891304](https://pubmed.ncbi.nlm.nih.gov/11891304/)
19. Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, Harris SR, 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 neglected tropical diseases*. 2012; 6(2):e1552. doi: [10.1371/journal.pntd.0001552](https://doi.org/10.1371/journal.pntd.0001552) PMID: [22389744](https://pubmed.ncbi.nlm.nih.gov/22389744/)
20. Traore B, Diarra B, Dembele BP, Somboro AM, Hammond AS, Siddiqui S, et al. Molecular strain typing of *Mycobacterium tuberculosis* complex in Bamako, Mali. *The international journal of tuberculosis and lung disease*. 2012; 16(7):911–6.
21. Garnier T, Eiglmeier K, Camus JC, Medina N, Mansoor H, Pryor M, et al. The complete genome sequence of *Mycobacterium bovis*. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100(13):7877–82. PMID: [12788972](https://pubmed.ncbi.nlm.nih.gov/12788972/)
22. Orduna P, Cevallos MA, de Leon SP, Arvizu A, Hernandez-Gonzalez IL, Mendoza-Hernandez G, et al. Genomic and proteomic analyses of *Mycobacterium bovis* BCG Mexico 1931 reveal a diverse immunogenic repertoire against tuberculosis infection. *BMC genomics*. 2011; 12:493. doi: [10.1186/1471-2164-12-493](https://doi.org/10.1186/1471-2164-12-493) PMID: [21981907](https://pubmed.ncbi.nlm.nih.gov/21981907/)
23. Brosch R, Gordon SV, Garnier T, Eiglmeier K, Frigui W, Valenti P, et al. Genome plasticity of BCG and impact on vaccine efficacy. *Proceedings of the National Academy of Sciences of the United States of America*. 2007; 104(13):5596–601. PMID: [17372194](https://pubmed.ncbi.nlm.nih.gov/17372194/)
24. Seki M, Honda I, Fujita I, Yano I, Yamamoto S, Koyama A. Whole genome sequence analysis of *Mycobacterium bovis* bacillus Calmette-Guerin (BCG) Tokyo 172: a comparative study of BCG vaccine substrains. *Vaccine*. 2009; 27(11):1710–6. doi: [10.1016/j.vaccine.2009.01.034](https://doi.org/10.1016/j.vaccine.2009.01.034) PMID: [19200449](https://pubmed.ncbi.nlm.nih.gov/19200449/)
25. Cohen KA, Abeel T, Manson McGuire A, Desjardins CA, Munsamy V, Shea TP, et al. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of *Mycobacterium tuberculosis* Isolates from KwaZulu-Natal. *PLoS Med*. 2015; 12(9):e1001880. doi: [10.1371/journal.pmed.1001880](https://doi.org/10.1371/journal.pmed.1001880) PMID: [26418737](https://pubmed.ncbi.nlm.nih.gov/26418737/)
26. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nature genetics*. 2013; 45(2):172–9. doi: [10.1038/ng.2517](https://doi.org/10.1038/ng.2517) PMID: [23291586](https://pubmed.ncbi.nlm.nih.gov/23291586/)
27. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, et al. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nature genetics*. 2013; 45(10):1255–60. doi: [10.1038/ng.2735](https://doi.org/10.1038/ng.2735) PMID: [23995137](https://pubmed.ncbi.nlm.nih.gov/23995137/)
28. Larsen MH, Biermann K, Tandberg S, Hsu T, Jacobs WR Jr. Genetic Manipulation of *Mycobacterium tuberculosis*. *Curr Protoc Microbiol*. 2007; Chapter 10:Unit 10A 2.
29. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS one*. 2014; 9(11):e112963. doi: [10.1371/journal.pone.0112963](https://doi.org/10.1371/journal.pone.0112963) PMID: [25409509](https://pubmed.ncbi.nlm.nih.gov/25409509/)
30. Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, et al. Finished bacterial genomes from shotgun sequence data. *Genome research*. 2012; 22(11):2270–7. doi: [10.1101/gr.141515.112](https://doi.org/10.1101/gr.141515.112) PMID: [22829535](https://pubmed.ncbi.nlm.nih.gov/22829535/)

31. Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome biology*. 2011; 12(1):R1. doi: [10.1186/gb-2011-12-1-r1](https://doi.org/10.1186/gb-2011-12-1-r1) PMID: [21205303](https://pubmed.ncbi.nlm.nih.gov/21205303/)
32. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome biology*. 2004; 5(2):R12. PMID: [14759262](https://pubmed.ncbi.nlm.nih.gov/14759262/)
33. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*. 2010; 11:119. Epub 2010/03/10. doi: [10.1186/1471-2105-11-119](https://doi.org/10.1186/1471-2105-11-119) PMID: [20211023](https://pubmed.ncbi.nlm.nih.gov/20211023/)
34. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*. 1997; 25(5):955–64. Epub 1997/03/01. PMID: [9023104](https://pubmed.ncbi.nlm.nih.gov/9023104/)
35. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*. 2007; 35(9):3100–8. Epub 2007/04/25. PMID: [17452365](https://pubmed.ncbi.nlm.nih.gov/17452365/)
36. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz HR, et al. The Pfam protein families database. *Nucleic Acids Res*. 2008; 36(Database issue):D281–8. PMID: [18039703](https://pubmed.ncbi.nlm.nih.gov/18039703/)
37. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, et al. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*. 2001; 29(1):41–3. Epub 2000/01/11. PMID: [11125044](https://pubmed.ncbi.nlm.nih.gov/11125044/)
38. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 1999; 27(1):29–34. Epub 1998/12/10. PMID: [9847135](https://pubmed.ncbi.nlm.nih.gov/9847135/)
39. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science*. 1997; 278(5338):631–7. Epub 1997/10/24. PMID: [9381173](https://pubmed.ncbi.nlm.nih.gov/9381173/)
40. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005; 21(18):3674–6. PMID: [16081474](https://pubmed.ncbi.nlm.nih.gov/16081474/)
41. Tian W, Arakaki AK, Skolnick J. EFICAZ: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res*. 2004; 32(21):6226–39. Epub 2004/12/04. PMID: [15576349](https://pubmed.ncbi.nlm.nih.gov/15576349/)
42. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 2011; 8(10):785–6. Epub 2011/10/01. doi: [10.1038/nmeth.1701](https://doi.org/10.1038/nmeth.1701) PMID: [21959131](https://pubmed.ncbi.nlm.nih.gov/21959131/)
43. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001; 305(3):567–80. Epub 2001/01/12. PMID: [11152613](https://pubmed.ncbi.nlm.nih.gov/11152613/)
44. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of clinical microbiology*. 1997; 35(4):907–14. PMID: [9157152](https://pubmed.ncbi.nlm.nih.gov/9157152/)
45. Demay C, Liens B, Burguiere T, Hill V, Couvin D, Millet J, et al. SITVITWEB—a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infect Genet Evol*. 2012; 12(4):755–66. doi: [10.1016/j.meegid.2012.02.004](https://doi.org/10.1016/j.meegid.2012.02.004) PMID: [22365971](https://pubmed.ncbi.nlm.nih.gov/22365971/)
46. Griggs A, Wapinski I, Wortman J, Haas B. SYNERGY2: Accurate and scalable ortholog identification. in preparation. 2014.
47. Wapinski I, Pfeffer A, Friedman N, Regev A. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*. 2007; 23(13):i549–58. Epub 2007/07/25. PMID: [17646342](https://pubmed.ncbi.nlm.nih.gov/17646342/)
48. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. *Nature*. 2007; 449(7158):54–61. Epub 2007/09/07. PMID: [17805289](https://pubmed.ncbi.nlm.nih.gov/17805289/)
49. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006; 22(21):2688–90. Epub 2006/08/25. PMID: [16928733](https://pubmed.ncbi.nlm.nih.gov/16928733/)
50. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(7):2567–72. Epub 2005/02/11. PMID: [15701695](https://pubmed.ncbi.nlm.nih.gov/15701695/)
51. Palmer KL, Godfrey P, Griggs A, Kos VN, Zucker J, Desjardins C, et al. Comparative genomics of enterococci: variation in *Enterococcus faecalis*, clade structure in *E. faecium*, and defining characteristics of *E. gallinarum* and *E. casseliflavus*. *mBio*. 2012; 3(1):e00318–11. Epub 2012/02/23. doi: [10.1128/mBio.00318-11](https://doi.org/10.1128/mBio.00318-11) PMID: [22354958](https://pubmed.ncbi.nlm.nih.gov/22354958/)
52. Wilgenbusch JC, Swofford D. Inferring evolutionary trees with PAUP\*. *Current protocols in bioinformatics / editorial board, Baxevanis Andreas D [et al]*. 2003;Chapter 6:Unit 6 4.



53. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25(14):1754–60. doi: [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324) PMID: [19451168](https://pubmed.ncbi.nlm.nih.gov/19451168/)
54. de Jong BC, Antonio M, Awine T, Ogungbemi K, de Jong YP, Gagneux S, et al. Use of spoligotyping and large sequence polymorphisms to study the population structure of the *Mycobacterium tuberculosis* complex in a cohort study of consecutive smear-positive tuberculosis cases in The Gambia. *Journal of clinical microbiology*. 2009; 47(4):994–1001. doi: [10.1128/JCM.01216-08](https://doi.org/10.1128/JCM.01216-08) PMID: [19193842](https://pubmed.ncbi.nlm.nih.gov/19193842/)
55. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS one*. 2010; 5(3):e9490. Epub 2010/03/13. doi: [10.1371/journal.pone.0009490](https://doi.org/10.1371/journal.pone.0009490) PMID: [20224823](https://pubmed.ncbi.nlm.nih.gov/20224823/)
56. Storey JD. A direct approach to false discovery rates. *J R Statist Soc B*. 2002; 64(Part 3):479–98.
57. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*. 2009; 4(7):1073–81. doi: [10.1038/nprot.2009.86](https://doi.org/10.1038/nprot.2009.86) PMID: [19561590](https://pubmed.ncbi.nlm.nih.gov/19561590/)
58. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC bioinformatics*. 2009; 10:296. doi: [10.1186/1471-2105-10-296](https://doi.org/10.1186/1471-2105-10-296) PMID: [19765293](https://pubmed.ncbi.nlm.nih.gov/19765293/)
59. Hain Genotype MTBDRsl. Available from: <http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/genotype-mtbdsl.html>
60. Hain Genotype MTBDRplus. Available from: <http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/genotype-mtbdplus.html>.
61. Smith NH, Gordon SV, de la Rúa-Domenech R, Clifton-Hadley RS, Hewinson RG. Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*. *Nature reviews Microbiology*. 2006; 4(9):670–81. PMID: [16912712](https://pubmed.ncbi.nlm.nih.gov/16912712/)
62. Koeck JL, Fabre M, Simon F, Daffe M, Garnotel E, Matan AB, et al. Clinical characteristics of the smooth tubercle bacilli '*Mycobacterium canettii*' infection suggest the existence of an environmental reservoir. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*. 2011; 17(7):1013–9.
63. Galtier N, Depaulis F, Barton NH. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics*. 2000; 155(2):981–7. PMID: [10835415](https://pubmed.ncbi.nlm.nih.gov/10835415/)
64. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *The Lancet Infectious diseases*. 2013; 13(2):137–46. doi: [10.1016/S1473-3099\(12\)70277-3](https://doi.org/10.1016/S1473-3099(12)70277-3) PMID: [23158499](https://pubmed.ncbi.nlm.nih.gov/23158499/)
65. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular microbiology*. 2003; 48(1):77–84. PMID: [12657046](https://pubmed.ncbi.nlm.nih.gov/12657046/)
66. Sassetti CM, Rubin EJ. Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100(22):12989–94. PMID: [14569030](https://pubmed.ncbi.nlm.nih.gov/14569030/)
67. Rengarajan J, Bloom BR, Rubin EJ. Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(23):8327–32. PMID: [15928073](https://pubmed.ncbi.nlm.nih.gov/15928073/)
68. Homolka S, Post E, Oberhauser B, George AG, Westman L, Dafaie F, et al. High genetic diversity among *Mycobacterium tuberculosis* complex strains from Sierra Leone. *BMC Microbiol*. 2008; 8:103. doi: [10.1186/1471-2180-8-103](https://doi.org/10.1186/1471-2180-8-103) PMID: [18578864](https://pubmed.ncbi.nlm.nih.gov/18578864/)
69. Hillemann D, Rusch-Gerdes S, Richter E. Feasibility of the GenoType MTBDRsl assay for fluoroquinolone, amikacin-capreomycin, and ethambutol resistance testing of *Mycobacterium tuberculosis* strains and clinical specimens. *Journal of clinical microbiology*. 2009; 47(6):1767–72. doi: [10.1128/JCM.00081-09](https://doi.org/10.1128/JCM.00081-09) PMID: [19386845](https://pubmed.ncbi.nlm.nih.gov/19386845/)
70. Ignatyeva O, Kontsevaya I, Kovalyov A, Balabanova Y, Nikolayevskyy V, Toit K, et al. Detection of resistance to second-line antituberculosis drugs by use of the genotype MTBDRsl assay: a multicenter evaluation and feasibility study. *Journal of clinical microbiology*. 2012; 50(5):1593–7. doi: [10.1128/JCM.00039-12](https://doi.org/10.1128/JCM.00039-12) PMID: [22378910](https://pubmed.ncbi.nlm.nih.gov/22378910/)
71. Kiet VS, Lan NT, An DD, Dung NH, Hoa DV, van Vinh Chau N, et al. Evaluation of the MTBDRsl test for detection of second-line-drug resistance in *Mycobacterium tuberculosis*. *Journal of clinical microbiology*. 2010; 48(8):2934–9. doi: [10.1128/JCM.00201-10](https://doi.org/10.1128/JCM.00201-10) PMID: [20573868](https://pubmed.ncbi.nlm.nih.gov/20573868/)
72. Kontsevaya I, Ignatyeva O, Nikolayevskyy V, Balabanova Y, Kovalyov A, Kritsky A, et al. Diagnostic accuracy of the genotype MTBDRsl assay for rapid diagnosis of extensively drug-resistant tuberculosis in HIV-coinfected patients. *Journal of clinical microbiology*. 2013; 51(1):243–8. doi: [10.1128/JCM.02513-12](https://doi.org/10.1128/JCM.02513-12) PMID: [23152552](https://pubmed.ncbi.nlm.nih.gov/23152552/)

73. Lacoma A, Garcia-Sierra N, Prat C, Maldonado J, Ruiz-Manzano J, Haba L, et al. GenoType MTBDRsl for molecular detection of second-line-drug and ethambutol resistance in *Mycobacterium tuberculosis* strains and clinical samples. *Journal of clinical microbiology*. 2012; 50(1):30–6. doi: [10.1128/JCM.05274-11](https://doi.org/10.1128/JCM.05274-11) PMID: [22075597](https://pubmed.ncbi.nlm.nih.gov/22075597/)
74. Orikiriza P, Tibenderana B, Siedner MJ, Mueller Y, Byarugaba F, Moore CC, et al. Low Resistance to First and Second Line Anti-Tuberculosis Drugs among Treatment Naive Pulmonary Tuberculosis Patients in Southwestern Uganda. *PloS one*. 2015; 10(2):e0118191. doi: [10.1371/journal.pone.0118191](https://doi.org/10.1371/journal.pone.0118191) PMID: [25658921](https://pubmed.ncbi.nlm.nih.gov/25658921/)
75. Singhal R, Myneedu VP, Arora J, Singh N, Bhalla M, Verma A, et al. Early detection of multi-drug resistance and common mutations in *Mycobacterium tuberculosis* isolates from Delhi using Geno-Type MTBDRplus assay. *Indian journal of medical microbiology*. 2015; 33 Suppl:S46–52.
76. Finken M, Kirschner P, Meier A, Wrede A, Bottger EC. Molecular basis of streptomycin resistance in *Mycobacterium tuberculosis*: alterations of the ribosomal protein S12 gene and point mutations within a functional 16S ribosomal RNA pseudoknot. *Molecular microbiology*. 1993; 9(6):1239–46. PMID: [7934937](https://pubmed.ncbi.nlm.nih.gov/7934937/)
77. Wong SY, Lee JS, Kwak HK, Via LE, Boshoff HI, Barry CE 3rd. Mutations in gidB confer low-level streptomycin resistance in *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*. 2011; 55(6):2515–22. doi: [10.1128/AAC.01814-10](https://doi.org/10.1128/AAC.01814-10) PMID: [21444711](https://pubmed.ncbi.nlm.nih.gov/21444711/)
78. Farooqi JQ, Khan E, Alam SM, Ali A, Hasan Z, Hasan R. Line probe assay for detection of rifampicin and isoniazid resistant tuberculosis in Pakistan. *JPMA The Journal of the Pakistan Medical Association*. 2012; 62(8):767–72. PMID: [23862246](https://pubmed.ncbi.nlm.nih.gov/23862246/)
79. WHO. Global Tuberculosis Report 2014. Available from: [http://apps.who.int/iris/bitstream/10665/137094/1/9789241564809\\_eng.pdf?ua=1](http://apps.who.int/iris/bitstream/10665/137094/1/9789241564809_eng.pdf?ua=1).
80. WHO. Tuberculosis country profiles: Mali. Available from: <http://www.who.int/tb/country/data/profiles/en/>.
81. McGuire AM, Weiner B, Park ST, Wapinski I, Raman S, Dolganov G, et al. Comparative analysis of *Mycobacterium* and related Actinomycetes yields insight into the evolution of *Mycobacterium tuberculosis* pathogenesis. *BMC genomics*. 2012; 13:120. doi: [10.1186/1471-2164-13-120](https://doi.org/10.1186/1471-2164-13-120) PMID: [22452820](https://pubmed.ncbi.nlm.nih.gov/22452820/)
82. Gopinath K, Moosa A, Mizrahi V, Warner DF. Vitamin B(12) metabolism in *Mycobacterium tuberculosis*. *Future microbiology*. 2013; 8(11):1405–18. doi: [10.2217/fmb.13.113](https://doi.org/10.2217/fmb.13.113) PMID: [24199800](https://pubmed.ncbi.nlm.nih.gov/24199800/)
83. Lavollay M, Arthur M, Fourgeaud M, Dubost L, Marie A, Veziris N, et al. The peptidoglycan of stationary-phase *Mycobacterium tuberculosis* predominantly contains cross-links generated by L,D-transpeptidation. *Journal of bacteriology*. 2008; 190(12):4360–6. doi: [10.1128/JB.00239-08](https://doi.org/10.1128/JB.00239-08) PMID: [18408028](https://pubmed.ncbi.nlm.nih.gov/18408028/)
84. Schoonmaker MK, Bishai WR, Lamichhane G. Nonclassical transpeptidases of *Mycobacterium tuberculosis* alter cell size, morphology, the cytosolic matrix, protein localization, virulence, and resistance to beta-lactams. *Journal of bacteriology*. 2014; 196(7):1394–402. doi: [10.1128/JB.01396-13](https://doi.org/10.1128/JB.01396-13) PMID: [24464457](https://pubmed.ncbi.nlm.nih.gov/24464457/)
85. Agarwal N, Lamichhane G, Gupta R, Nolan S, Bishai WR. Cyclic AMP intoxication of macrophages by a *Mycobacterium tuberculosis* adenylate cyclase. *Nature*. 2009; 460(7251):98–102. doi: [10.1038/nature08123](https://doi.org/10.1038/nature08123) PMID: [19516256](https://pubmed.ncbi.nlm.nih.gov/19516256/)
86. Lu P, Lill H, Bald D. ATP synthase in mycobacteria: special features and implications for a function as drug target. *Biochimica et biophysica acta*. 2014; 1837(7):1208–18. doi: [10.1016/j.bbabi.2014.01.022](https://doi.org/10.1016/j.bbabi.2014.01.022) PMID: [24513197](https://pubmed.ncbi.nlm.nih.gov/24513197/)
87. Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, et al. A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science*. 2005; 307(5707):223–7. PMID: [15591164](https://pubmed.ncbi.nlm.nih.gov/15591164/)
88. Gioffre A, Infante E, Aguilar D, Santangelo MP, Klepp L, Amadio A, et al. Mutation in mce operons attenuates *Mycobacterium tuberculosis* virulence. *Microbes and infection / Institut Pasteur*. 2005; 7(3):325–34.
89. Thorel MF. Isolation of *Mycobacterium africanum* from monkeys. *Tubercle*. 1980; 61(2):101–4. PMID: [6776671](https://pubmed.ncbi.nlm.nih.gov/6776671/)
90. Thorel MF. [Mycobacteria identified in a centre for veterinary research between 1973 and 1979 (author's transl)]. *Annales de microbiologie*. 1980; 131(1):61–9. PMID: [6767431](https://pubmed.ncbi.nlm.nih.gov/6767431/)
91. Coscolla M, Lewin A, Metzger S, Maetz-Rennsing K, Calvignac-Spencer S, Nitsche A, et al. Novel *Mycobacterium tuberculosis* complex isolate from a wild chimpanzee. *Emerging infectious diseases*. 2013; 19(6):969–76. doi: [10.3201/eid1906.121012](https://doi.org/10.3201/eid1906.121012) PMID: [23735084](https://pubmed.ncbi.nlm.nih.gov/23735084/)

92. Rahim Z, Mollers M, te Koppele-Vije A, de Beer J, Zaman K, Matin MA, et al. Characterization of *Mycobacterium africanum* subtype I among cows in a dairy farm in Bangladesh using spoligotyping. *The Southeast Asian journal of tropical medicine and public health*. 2007; 38(4):706–13. PMID: [17883011](#)
93. Alfredeisen S, Saxegaard F. An outbreak of tuberculosis in pigs and cattle caused by *Mycobacterium africanum*. *The Veterinary record*. 1992; 131(3):51–3. PMID: [1441162](#)
94. Gudan A, Artukovic B, Cvetnic Z, Spicic S, Beck A, Hohsteter M, et al. Disseminated tuberculosis in hyrax (*Procavia capensis*) caused by *Mycobacterium africanum*. *Journal of zoo and wildlife medicine: official publication of the American Association of Zoo Veterinarians*. 2008; 39(3):386–91.
95. Michel AL, Muller B, van Helden PD. *Mycobacterium bovis* at the animal-human interface: a problem, or not? *Veterinary microbiology*. 2010; 140(3–4):371–81. doi: [10.1016/j.vetmic.2009.08.029](#) PMID: [19773134](#)
96. Mandal S, Bradshaw L, Anderson LF, Brown T, Evans JT, Drobniowski F, et al. Investigating transmission of *Mycobacterium bovis* in the United Kingdom in 2005 to 2008. *Journal of clinical microbiology*. 2011; 49(5):1943–50. doi: [10.1128/JCM.02299-10](#) PMID: [21430093](#)
97. Gonzalo-Asensio J, Malaga W, Pawlik A, Astarie-Dequeker C, Passemar C, Moreau F, et al. Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111(31):11491–6. doi: [10.1073/pnas.1406693111](#) PMID: [25049399](#)
98. Simeone R, Bottai D, Brosch R. ESX/type VII secretion systems and their role in host-pathogen interaction. *Current opinion in microbiology*. 2009; 12(1):4–10. doi: [10.1016/j.mib.2008.11.003](#) PMID: [19155186](#)
99. Houben EN, Korotkov KV, Bitter W. Take five—Type VII secretion systems of Mycobacteria. *Biochimica et biophysica acta*. 2014; 1843(8):1707–16. doi: [10.1016/j.bbamcr.2013.11.003](#) PMID: [24263244](#)
100. Pym AS, Brodin P, Brosch R, Huerre M, Cole ST. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti*. *Molecular microbiology*. 2002; 46(3):709–17. PMID: [12410828](#)
101. Pym AS, Brodin P, Majlessi L, Brosch R, Demangel C, Williams A, et al. Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. *Nature medicine*. 2003; 9(5):533–9. PMID: [12692540](#)
102. Oussalah A, Besseau C, Chery C, Jeannesson E, Gueant-Rodriguez RM, Anello G, et al. *Helicobacter pylori* serologic status has no influence on the association between fucosyltransferase 2 polymorphism (FUT2 461 G->A) and vitamin B-12 in Europe and West Africa. *The American journal of clinical nutrition*. 2012; 95(2):514–21. doi: [10.3945/ajcn.111.016410](#) PMID: [22237057](#)
103. Gueant-Rodriguez RM, Gueant JL, Debarb R, Thirion S, Hong LX, Bronowicki JP, et al. Prevalence of methylenetetrahydrofolate reductase 677T and 1298C alleles and folate status: a comparative study in Mexican, West African, and European populations. *The American journal of clinical nutrition*. 2006; 83(3):701–7. PMID: [16522920](#)
104. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *The Journal of biological chemistry*. 2003; 278(42):41148–59. PMID: [12869542](#)
105. Gopinath K, Venclovas C, Ioerger TR, Sacchettini JC, McKinney JD, Mizrahi V, et al. A vitamin B(1)(2) transporter in *Mycobacterium tuberculosis*. *Open biology*. 2013; 3(2):120175. doi: [10.1098/rsob.120175](#) PMID: [23407640](#)
106. Gehrre F, Antonio M, Otu JK, Sallah N, Secka O, Faal T, et al. Immunogenic *Mycobacterium africanum* strains associated with ongoing transmission in The Gambia. *Emerging infectious diseases*. 2013; 19(10):1598–604. doi: [10.3201/eid1910.121023](#) PMID: [24050158](#)
107. Niemann S, Rusch-Gerdes S, Joloba ML, Whalen CC, Guwatudde D, Ellner JJ, et al. *Mycobacterium africanum* subtype II is associated with two distinct genotypes and is a major cause of human tuberculosis in Kampala, Uganda. *Journal of clinical microbiology*. 2002; 40(9):3398–405. PMID: [12202584](#)
108. Ahmad S, Akbar PK, Wiker HG, Harboe M, Mustafa AS. Cloning, expression and immunological reactivity of two mammalian cell entry proteins encoded by the *mce1* operon of *Mycobacterium tuberculosis*. *Scandinavian journal of immunology*. 1999; 50(5):510–8. PMID: [10564554](#)
109. Bai G, Knapp GS, McDonough KA. Cyclic AMP signalling in mycobacteria: redirecting the conversation with a common currency. *Cellular microbiology*. 2011; 13(3):349–58. doi: [10.1111/j.1462-5822.2010.01562.x](#) PMID: [21199259](#)
110. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*. 1998; 393(6685):537–44. PMID: [9634230](#)

111. Banu S, Honore N, Saint-Joanis B, Philpott D, Prevost MC, Cole ST. Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Molecular microbiology*. 2002; 44(1):9–19. PMID: [11967065](#)
112. Godreuil S, Torrea G, Terru D, Chevenet F, Diagbouga S, Supply P, et al. First molecular epidemiology study of *Mycobacterium tuberculosis* in Burkina Faso. *Journal of clinical microbiology*. 2007; 45(3):921–7. PMID: [17251410](#)
113. Ani A, Bruvik T, Okoh Y, Agaba P, Agbaji O, Idoko J, et al. Genetic diversity of *Mycobacterium tuberculosis* Complex in Jos, Nigeria. *BMC Infect Dis*. 2010; 10:189. doi: [10.1186/1471-2334-10-189](#) PMID: [20579382](#)
114. Cadmus S, Hill V, van Soolingen D, Rastogi N. Spoligotype profile of *Mycobacterium tuberculosis* complex strains from HIV-positive and -negative patients in Nigeria: a comparative analysis. *Journal of clinical microbiology*. 2011; 49(1):220–6. doi: [10.1128/JCM.01241-10](#) PMID: [21048016](#)
115. Asiimwe BB, Ghebremichael S, Kallenius G, Koivula T, Joloba ML. *Mycobacterium tuberculosis* spoligotypes and drug susceptibility pattern of isolates from tuberculosis patients in peri-urban Kampala, Uganda. *BMC Infect Dis*. 2008; 8:101. doi: [10.1186/1471-2334-8-101](#) PMID: [18662405](#)