Diversity of *Mycobacterium tuberculosis* across evolutionary scales

Mary B. O'Neill[1,2], Tatum D. Mortimer[2], and Caitlin S. Pepperell[2§]

[1]Laboratory of Genetics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

[2]Departments of Medicine and Medical Microbiology and Immunology, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

[§]Corresponding author

Email addresses:

MBO: mrood@wisc.edu

TDM: tmortimer@wisc.edu

CSP: cspepper@medicine.wisc.edu

Abstract

Tuberculosis (TB) is a global public health emergency. Increasingly drug resistant strains of *Mycobacterium tuberculosis* (*M.tb*) continue to emerge and spread, highlighting the adaptability of this pathogen. Most studies of *M.tb* evolution have relied on 'between-host' samples, in which each person with TB is represented by a single *M.tb* isolate. However, individuals with TB commonly harbor populations of *M.tb* numbering in the billions. Here, we use analyses of *M.tb* diversity found within and between hosts to gain insight into the adaptation of this pathogen. We find that the amount of *M.tb* genetic diversity harbored by individuals with TB is similar to that of global between-host surveys of TB patients. This suggests that *M.tb* genetic diversity is generated within hosts and then lost as the infection is transmitted. In examining genomic data from *M.tb* samples within and between hosts with TB, we find that genes involved in the regulation, synthesis, and transportation of immunomodulatory cell envelope lipids appear repeatedly in the extremes of various statistical measures of diversity. Polyketide synthase and Mycobacterial membrane protein Large (mmpL) genes are particularly notable in this regard. In addition, we observe identical mutations emerging across samples from different TB patients. Taken together, our observations suggest that *M.tb* cell envelope lipids are targets of selection within hosts. These lipids are specific to pathogenic mycobacteria and, in some cases, human-pathogenic mycobacteria. We speculate that rapid adaptation of cell envelope lipids is facilitated by functional redundancy, flexibility in their metabolism, and their roles mediating interactions with the host.

Keywords

Pathogen evolution, adaptation, population genetics, natural variation, lipid metabolism

Author Summary

Tuberculosis (TB) is a grave threat to global public health, and is the second leading cause of death due to infectious disease. The causative agent, *Mycobacterium tuberculosis* (*M.tb*), has emerged in increasingly drug resistant forms that hamper our efforts to control TB. We need a better understanding of *M.tb* adaptation to guide development of more effective TB treatment and control strategies. The goal of this study was to gain insight into *M.tb* evolution within individual patients with TB. We found that TB patients harbor a surprisingly diverse population of *M.tb*. We further found evidence to suggest that the bacterial population evolves measurably in response to selection pressures imposed by the environment within hosts. Changes were particularly in notable in *M.tb* genes involved in the regulation, synthesis and transportation of lipids and glycolipids of the bacterial cell envelope. These findings have important implications for drug and vaccine development, and provide insight into TB host pathogen interactions.

Background

Mycobacterium tuberculosis (*M.tb*) causes over nine million new cases of tuberculosis (TB) per year, and is estimated to infect one-third of the world's population [1]. The emergence of increasingly drug resistant strains of *M.tb* [2] demonstrates the bacterium's ability to adapt to antibiotic pressures, despite limited genetic diversity [3]. Prior research has identified the influence of bottlenecks, population sub-division, and purifying selection on genetic diversity of *M.tb* circulating among human hosts [4–7]. In these studies, each TB patient was represented by a single *M.tb* strain isolated in pure culture. However, individuals with TB harbor a large

population of *M.tb* cells for a period of months to years, which raises the possibility of significant diversification of bacterial populations over the course of individual infections.

There are few studies of within-host evolution of *M.tb*. One example is a study of the transposable element IS5110 marker that found multiple lines of evidence suggestive of positive selection on *M.tb* populations within hosts [8]. Advances in sequencing technologies have since enabled detailed, genome-wide studies of the evolution of intra-host populations of both pathogenic and commensal microbes [9–14]. Whole-genome sequencing (WGS) studies of natural populations of *M.tb* have focused primarily on the emergence of drug resistance [14–17]. Here, we use analyses of genetic diversity in *M.tb* populations from TB patients' sputum samples to characterize the within-host evolution of *M.tb*, and compare patterns of diversity to global surveys of the bacterium.

Results and Discussion

Table 1. Within-host samples of *Mycobacterium tuberculosis*. Sample dates and resistance profiles are from [14]; sample timing is in reference to treatment initiation. ABBREVIATIONS: DOC – Depth of Coverage per Site. INH – Isoniazid, SM – Streptomycin, RMP – Rifampicin, ETH – Ethambutol

			Pre Removal Poorly Mapped Regions		Post Removal Poorly Mapped Regions		Coverage ≥ 50X
Sputum Sample	**Sample Date**	**Resistance Profile**	**Mean DOC**	**Percent Genome**	**Mean DOC**	**Percent Genome**	**Percent Genome**
a1	-7th day	sensitive	97	98.1	81	88.1	74.7
a2	19th month	INH	103	98.3	89	88.1	84.3
a3	24th month	INH	312	98.3	254	88.1	87.9
b1	18th month	INH, SM	38	97.1	32	88.2	< 2
b2	35th month	INH, SM, RMP	159	98.5	139	88.2	87.7
c1	-24th day	INH, SM, RMP, ETH	147	98.1	127	88.2	87.1
c2	11th month	INH, SM, RMP, ETH	128	98.1	109	88.2	85.1

Genome-wide variation

We quantified genetic diversity of within-host populations of *M.tb* (Table 1) using two estimators: nucleotide diversity (π) and Watterson's theta (θ_W). These statistics were calculated

with PoPoolation software, which accounts for the effects of sequencing errors on low frequency variant allele calls in pooled sequence data [18,19]. Average measures of diversity and temporal trends in diversity varied among the three patients. Data from all patients showed effects of linkage in the relatively uniform behavior of statistics across the genome (Fig. 1). This is consistent with the clonal reproduction of *M.tb* [20].

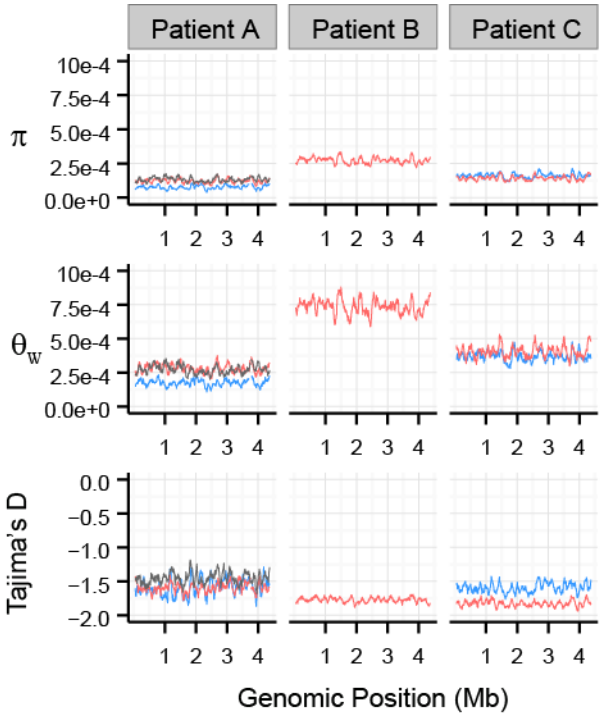


Fig. 1. Patterns of nucleotide diversity (π), Watterson's theta (□), and Tajima's D across the *Mycobacterium tuberculosis* chromosome. Sliding-window analyses were performed using 100-Kb windows with a step-size of 10-Kb on a uniformly subsampled alignment for each sample (50X sequence coverage). Chromosomal coordinates reflect the genomic positions of the reference strain H37Rv, against which pooled-sequence reads were mapped. Blue lines correspond to the first temporal sample of each patient; coral lines correspond to the second temporal sample of each patient; the dark gray line corresponds to the third temporal sample of patient A. Tajima's D is negative for all patients and time points, consistent with population expansion, purifying selection, and/or recent selective sweep(s). All three statistics are relatively even across the sliding windows, consistent with linkage of sites in the *M.tb* genome. (Refer to S1 Figure for a comparison of samples from patient B.)

There was a striking increase in diversity in samples from patient B: diversity estimates increased by an order of magnitude between sample b1, taken 18 months into treatment, and sample b2, taken 35 months into treatment (Table 2, S1 Figure). This difference is not

attributable to sequencing coverage since calculations were performed on uniformly sub-sampled data (*see methods*). A pre-treatment sputum sample from patient B was not available, and at 18 months into treatment (sample b1) resistance to isoniazid (INH) and streptomycin (SM) were both detected (Table 1) [14]. The extremely low diversity of this sample (Table 2) could be due to a recent clonal replacement event (i.e. driven by drug resistance mutations), in which all variation linked to the site of selection fixes along with the beneficial mutation. By 35 months we observe a recovery of diversity, with both π and θ_W increasing by nearly an order of magnitude (Table 2). Diversity of this sample is comparable to samples from the other patients in the study.

Table 2. Nucleotide diversity (π) and Watterson's theta (θ_W) estimates from *Mycobacterium tuberculosis* populations. Global and regional datasets are described in Supplementary Table S1. Estimators for global and regional datasets were calculated on genome-wide alignments using EggLib software [21]. Estimators were calculated on sub-sampled datasets (COV – coverage of sub-sampling) for pooled populations (a-c) using PoPoolation software [18].

Population		**π**	**θ_W**
Between-host	*n*		
Global	201	3.55×10^{-4}	1.50×10^{-3}
Lineage 2	37	1.04×10^{-4}	2.80×10^{-4}
EAS	18	2.30×10^{-4}	3.53×10^{-4}
China	14	2.52×10^{-4}	3.20×10^{-4}
Within-host	COV		
a1	50X	7.57×10^{-5}	1.74×10^{-4}
a2	50X	1.22×10^{-4}	2.81×10^{-4}
a3	50X	1.32×10^{-4}	2.74×10^{-4}
b1	25X	2.71×10^{-5}	4.32×10^{-5}
b2	25X	2.01×10^{-4}	3.51×10^{-4}
b2	50X	2.70×10^{-4}	7.29×10^{-4}
c1	50X	1.58×10^{-4}	3.54×10^{-4}
c2	50X	1.45×10^{-4}	4.14×10^{-4}

Measures of genome-wide diversity (π and θ_W) from *M.tb* populations within individual hosts are similar to estimates from regionally extant, between-host populations of bacteria (Table 2). Analysis of variable number tandem repeat genotypes from the within-host samples suggest

each population was initiated by a single clone [14]: these do not appear to be mixed infections, which have been documented [22]. Similar measures of diversity from within- and between-host populations of *M.tb* suggest that genetic diversity is repeatedly generated *in situ* within hosts and subsequently lost during transmission.

M.tb infections are initiated by droplets containing as few as 1-3 bacilli [23,24]; this infectious inoculum subsequently undergoes massive expansion within the host, eventually reaching a census population size of 10^7-10^{10} cells [5]. It is likely that these transmission bottlenecks are responsible for much of the loss of *M.tb* diversity generated within individual hosts. The ecology of TB may also contribute to loss of *M.tb* diversity during transmission. Discontinuities in TB transmission networks result in fine-scale sub-division of *M.tb* populations [5]. Founder effects traceable to historical environmental shifts favoring epidemic TB have also been observed in *M.tb* populations [5,6]. Purifying selection at the between-host level [4,25] is another potential contributor to loss of *M.tb* genetic diversity generated within hosts.

While estimates of π are similar among all population comparisons of *M.tb*, θ_W is higher for the global population than for within-host and regional populations. This could be due to shallow sampling across multiple *M.tb* sub-populations: alleles that are private to regional populations, but relatively common within them, would have disproportionate effects on segregating sites in global versus regional surveys.

Departure from expectations under a neutral model: genome-wide and gene by gene Tajima's D

We used Tajima's D [19,25] to assess whether within-host populations of *M.tb* depart from assumptions of neutral models of evolution. Tajima's D was negative and relatively even across the genome in all within-host samples, consistent with linkage of sites (Fig. 1). Negative

values of Tajima's D indicate an excess of rare variants relative to expectations for a neutrally evolving population of constant size. Excess rare variants can result from population expansion, purifying selection, or recent selective sweeps [25].

The demographic history of a sample will affect all sites in a fully linked genome, whereas effects of natural selection may be observable in different patterns of variation among loci associated with different functions. In order to investigate differences in the influence of selection on genes with distinct functions, we calculated Tajima's D for individual genes. We compared relative values of the statistic in each within-host sample, where all genes have the same demographic history. Using this approach, we found that the functional categories "lipid metabolism" (LIP) and "conserved hypotheticals" (CHP) are significantly enriched in the bottom (i.e. more negative) 5% tail of the distribution of gene-wise Tajima's D values of all independent patient samples (Fig. 2). Drug resistance associated "targets of independent mutation" (TIM), "secondary metabolite biosynthesis, transport, and catabolism" (COG:Q), and "replication, recombination and repair" (COG:L) are significantly enriched in the bottom (i.e. more negative) 5% tail of the distribution of gene-wise Tajima's D values of most within-host *M.tb* population samples (Fig. 2).

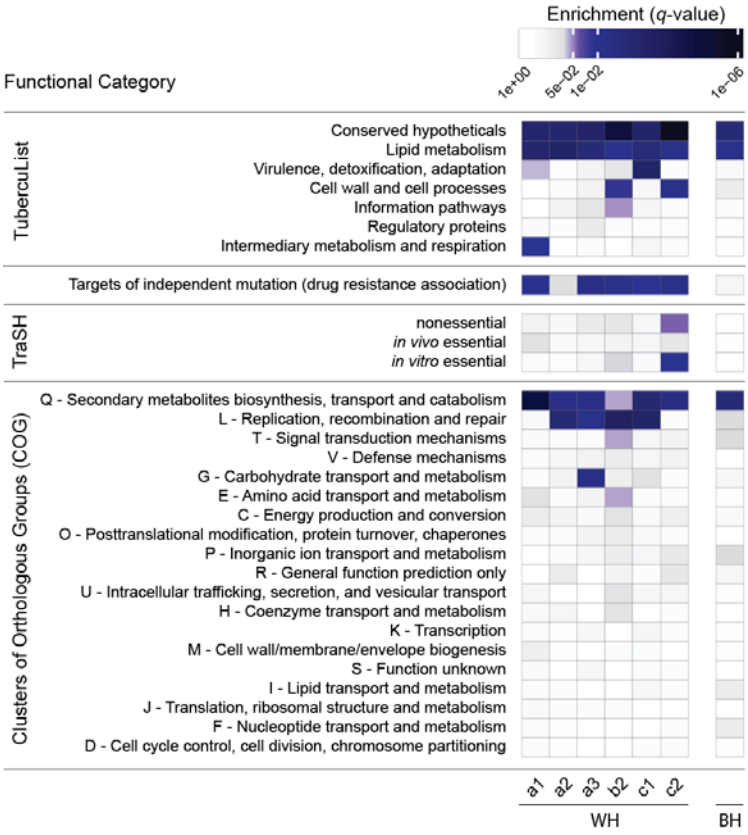


Fig. 2. Enrichment of annotation categories among genes with extreme negative values of Tajima's D. Within each sample (labeled at the bottom of the heatmap), gene-wise Tajima's D values were compared and the bottom 5% of genes in the distribution were tested for overrepresentation of functional categories using a two-sided Fisher's exact test. To account for multiple hypothesis testing, a false discovery rate of 5% was used and the resulting *q*-values are displayed. The manually curated TubercuList "conserved hypotheticals" and "lipid metabolism" categories [27], as well as the computationally predicted COG:Q "secondary metabolites biosynthesis, transport and catabolism" [28] are notable for their consistent enrichment at both the within- and between-host scales. WH – within-host; BH – between-host.

We performed a similar analysis in a between-host sample of 201 *M.tb* isolates collected around the globe (S2 Table contains descriptions of the strains) [29]. Between-host, gene-wise values of Tajima's D were strongly negative (Fig. 3). Recent global population expansion and background selection are likely contributors to genome-wide patterns of Tajima's D at this scale [4,5]. Consistent with findings for within-host populations of *M.tb*, LIP, CHP, and COG:Q categories were significantly enriched in the bottom 5% tail of the between-host distribution of Tajima's D (Fig. 2). This suggests that adaptive pressures leading to excess rare variants in these genes are stable across a range of human genetic backgrounds and environments. Expression of

genes in the COG:Q category has been shown to vary over the course of TB infections [15], which is consistent with their gene products being targets of selection within hosts. Interestingly, LIP and CHP were also enriched in the top 5% of between-host Tajima's D values (S3 Table). The enrichment of LIP and CHP in both tails suggests that distinct genes within these categories are subject to different regimes of selection.

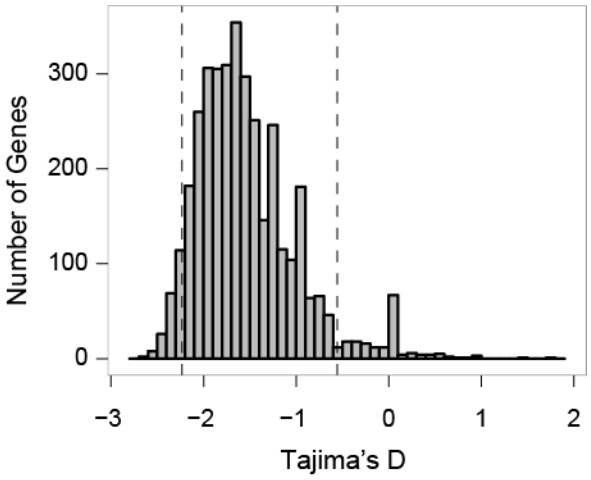


Fig. 3. Distribution of gene-wise Tajima's D values from between-host isolates. Tajima's D was calculated for every gene annotated in the H37Rv reference strain (that was resolved at ≥ 55% in ≥ 75% of strains) from an alignment of 201 globally extant strains of *Mycobacterium tuberculosis*. This between-host dataset is representative of all major lineages described to date (*see S2 Table for a description of strains*). Dotted lines delineate the 5th and 95th percentiles of the distribution.

Low values of within- and between-host Tajima's D in COG:Q (234 genes in category) are driven in part by genes that are also categorized as LIP (272 genes in category). Eighty-eight genes overlap between the two classification schemes, of which 44 are in the bottom tail of at least one within-host sample, and 18 are in the bottom tail of the between-host, gene-wise Tajima's D distribution. Among non-overlapping COG:Q genes, the mammalian cell entry (*mce*) family appears repeatedly in the bottom tail of gene-wise Tajima's D distributions (S4 Table). The *M.tb* genome encodes four *mce* gene clusters (*mce1-4*) [30]. Genes of the *mce1* and *mce4* loci are important for growth in a murine TB model [31], and the *mce1* operon has been implicated in modulating the host immune response [32,33].

Mechanisms of antibiotic resistance in *M.tb* are incompletely understood, and only a handful of genetic mutations have been experimentally demonstrated to confer resistance [34,35]. TIMs were recently identified in a global sample of *M.tb* strains: mutations at these loci were consistently identified in drug resistant strains of *M.tb* across distinct genetic backgrounds [35]. The TIM category was overrepresented in the lower tails of Tajima's D in multiple within-host samples (Fig. 2). This pattern was not replicated at the between-host scale. The within-host data are from individuals with drug resistant TB, suggesting that low values of Tajima's D are the residua of selective sweeps. The drug susceptibility profiles of between-host isolates were not reported by Comas *et al.* (2013); if the isolates were drug susceptible, we would not expect to see the effects of sweeps at TIM loci in the between-host sample. This would explain why these loci were not enriched in the bottom tail of the between-host Tajima's D.

Genes with high functional diversity: π_N/π_S outliers

In order to identify *M.tb* genes under positive or diversifying selection within hosts, we compared counts of non-synonymous changes per non-synonymous site (π_N) to counts of synonymous changes per synonymous site (π_S) in each gene. Stably maintained amino acid polymorphisms ($\pi_N/\pi_S > 1$) are indicative of diversifying selection or local sweeps (positive selection under a regime of restricted migration).

A high proportion of genes with positive values of π_N/π_S in the within-host samples are involved in lipid metabolism, particularly the synthesis of mycolic acids. The mycolic acid biosynthesis superpathway is over-represented among genes in the top 1% of values of π_N/π_S (p-value = 0.002, genes pooled across within-host samples) and among genes with $\pi_N/\pi_S > 1$ in at least one *M.tb* sample from each of the three patients (p-value = 0.04). An interesting example is

fatty-acid synthase (*fas*, *Rv2524c*), which is in the top 5% of π_N/π_S values in at least one sample from each of the three patients. Four SNPs that emerged in this gene were found in all three patients' *M.tb* populations. These findings suggest that the mycolic acid superpathway is remodeled over the course of individual TB infections.

Positive selection: genes with extreme measures of population differentiation between samples

Natural selection can lead to population differentiation when the relative fitness of genotypes varies among environments; empirical outlier analyses for loci with extreme measures of population differentiation are commonly used to identify candidates of positive selection [37–39]. Treating each serial sputum sample from an individual patient as a distinct *M.tb* population, we calculated pairwise F_{ST} values for all polymorphic sites (Fig. 4). In order to reduce biases introduced by variable coverage, we estimated F_{ST} from the uniformly sub-sampled sequencing data. While this approach alleviates coverage biases, it is likely to miss some potentially adaptive alleles that were sequenced at low coverage.

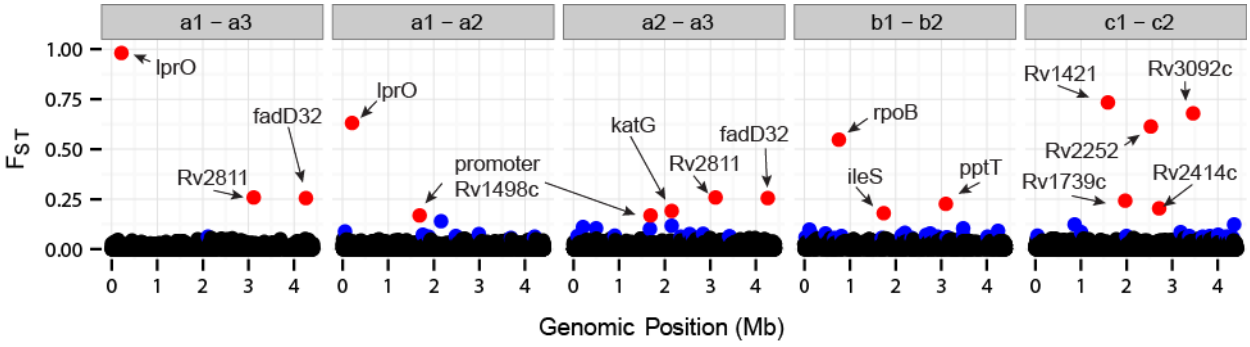


Fig. 4. Pairwise F_{ST} of polymorphic sites. Each temporal sample of a patient was treated as a population and F_{ST} was calculated for each variable site in the genome across the populations compared. Calculations were performed on merged uniform sub-samplings of the sequence data to reduce coverage biases (*see methods*). For Patients B and C, only one comparison was made as there are only two temporal samples; for Patient A, F_{ST} was calculated for each pairwise population comparison (a1-a2, a1-a3, a2-a3). Dots reflect single nucleotide polymorphisms (SNPs) across the H37Rv genome: dots in red represent SNPs in the top 0.1% of the F_{ST} distribution, and dots in blue represent SNPs in the top 1% of the F_{ST} distribution. Outliers of the F_{ST} distribution are likely to be under positive selection.

Our empirical outlier analysis identified extreme measures of population differentiation in well-characterized loci that mediate drug resistance (e.g. *katG* and *rpoB*), demonstrating the suitability of the method for identifying targets of positive selection within hosts. In addition to known drug resistance loci previously identified by Sun *et al.* (2012), we identified variants at several other loci that appeared to be under positive selection. LIP, CHP, COG:Q, and TIM annotations are prominent among genes harboring an outlier single nucleotide polymorphism (SNP). Ten of the 14 genes in the top 0.1% of each patient's F_{ST} distribution (collectively) belong to one or more of these categories (S5 Table).

An interesting example is *fadD32*, a gene in which we identified an outlier SNP in the *M.tb* population from patient A. FadD32 is essential for the synthesis of mycolic acids [40] and is currently being investigated as a target for new TB treatments [41,42]. Expression of *fadD32* was recently found to decrease over the course of infection in a patient who developed extensively drug-resistant TB [15]. In addition to identifying this gene in our outlier F_{ST} analysis, we found that it was in the 95th and 94th percentiles of between-host, gene-wise values of Tajima's D and π_N/π_S, respectively. In other words, this gene appears to be a target of selection within hosts, and is one of the most diverse genes in the *M.tb* genome. Despite recent pre-clinical promise of FadD32 as a target of coumarin compounds [41,42], our finding of high non-synonymous diversity in this gene suggests that the genetic barrier to acquisition of resistance at this locus may be low.

The gene encoding phosphopantetheinyl transferase (*pptT*) harbored an outlier SNP in the *M.tb* population of patient B (Fig. 4). PptT belongs to the functional categories LIP and COG:Q (S5 Table). PptT activates Pks13, a type-I polyketide synthase involved in the final step of mycolic acid biosynthesis, and various type-I polyketide synthases required for the synthesis of

lipid virulence factors [43,44]. Active FadD32 and Pks13 are involved in the final steps of mycolic acid condensation *in vitro* [40]; our detection of F_{ST} outliers in *fadD32* (patient A) and *pptT* (patient B) provides further support for the idea that *M.tb* mycolic acids are modulated over the course of individual TB infections.

Integrating across analyses and evolutionary scales: correlations among statistics

As described above, there are interesting congruencies in patterns of genetic variation across evolutionary scales: LIP, COG:Q, and CHP functional categories are enriched among genes with extremely low values of Tajima's D at both within- and between-host scales. Although these patterns of enrichment are consistent, their interpretation is not straightforward. Low values of Tajima's D can be observed among loci under negative (purifying) selection, as well as following selective sweeps (positive selection). For clonally reproducing organisms like *M.tb* [20], all variation linked to adaptive mutations should sweep to fixation - one would not expect to observe significant differences among loci in values of Tajima's D driven by positive selection. Nonetheless, specific functional categories appear consistently in the extremes of this statistic. In comparing different statistical measures of genetic variation, we also find indications that low TD is driven by positive selection on *M.tb* in the environment within hosts.

The within-host samples of *M.tb* are from individuals with drug resistant TB, in which resistance mutations have been positively selected. Our finding of an enrichment of loci mediating drug resistance among genes with extremely low Tajima's D (Fig. 2) supports its association with positive selection within hosts. Further supporting this link, we found that functional categories associated with low values of Tajima's D (i.e. LIP, COG:Q, and CHP) were prominent among genes harboring SNPs with outlier F_{ST} values (Fig. 4, S5 Table). One sub-

288 category of LIP genes – those involved in mycolic acid biosynthesis – was also prominent
289 among genes with high π_N/π_S in within-host samples.

290 In order to further clarify the interpretation of these patterns of variation, we integrated
291 analyses of *M.tb* genomic data across within- and between-host scales. We sought to clarify
292 whether or not the same genes within categories were driving the signal of low Tajima's D at the
293 within- versus between-host scale. Another goal was to identify genes with the strongest and
294 most consistent signatures of selection across evolutionary scales.

295 Using estimates of Tajima's D and π_N/π_S for 3,541 genes from within- and between-host
296 samples (S4 Table), we calculated correlations among statistics and identified genes with
297 extreme patterns of variation at both scales (Fig. 5). Gene-wise estimates of Tajima's D at the
298 within- and between-host scales were strongly correlated (Fig. 5C, correlation coefficient = 0.41,
299 *p*-value = 0.0). This suggests that many of the same genes are driving enrichment of LIP,
300 COG:Q, and CHP categories among low values of Tajima's D across evolutionary scales.
301 Within-host, gene-wise values of Tajima's D and π_N/π_S were negatively correlated (Fig. 5D,
302 correlation coefficient = -0.30, *p*-value = 4.0×10^{-37}): this suggests that low values of Tajima's D
303 within-hosts (and by extension between-hosts) may be driven by positive, rather than purifying,
304 selection within hosts. Further supporting a lack of association between low Tajima's D and
305 purifying selection, there was no overlap between genes in the 5th percentiles of gene-wise π_N/π_S
306 and Tajima's D at the between-host scale, nor at the within-host scale (Fig. 5B, 5D). Between-
307 host values of gene-wise Tajima's D were also negatively correlated with within-host values of
308 gene-wise π_N/π_S, but less strongly (correlation coefficient -0.14, *p*-value = 1.2×10^{-8}). This could
309 result from modification of within-host patterns of non-synonymous variation by neutral and/or
310 selective influences imposed by transmission. Between-host values of gene-wise Tajima's D

were not correlated with between-host values of gene-wise π_N/π_S (correlation coefficient -0.02, *p*-value = 0.34), suggesting that low values of Tajima's D observed between hosts is driven by events within hosts.

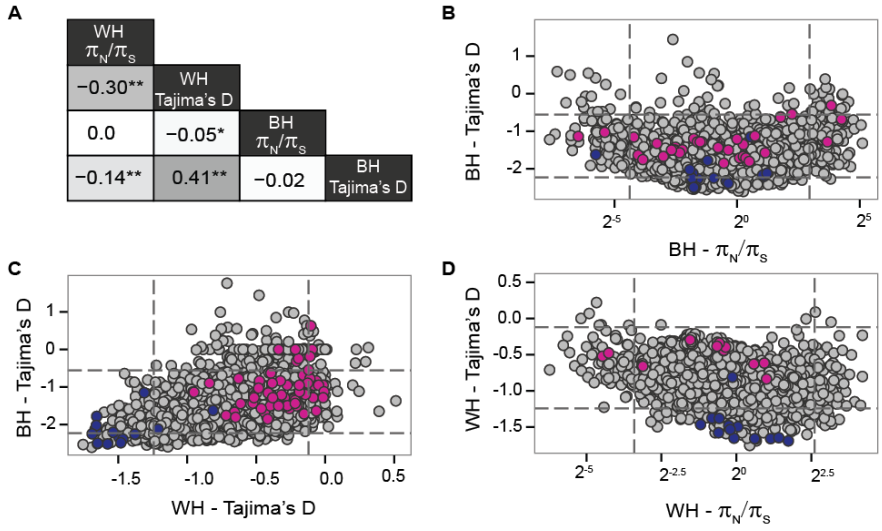


Fig. 5. Correlations among population genetic parameters across evolutionary scales. (A) Pearson's correlation coefficients. *Indicates significance at < 0.05 level; **Indicates significance at << 0.01 level. (B) Correlation between gene-wise estimates of Tajima's D and $\pi_N/\pi S$ at the between-host scale. (C) Correlation between gene-wise estimates of Tajima's D at the within- and between-host scales. (D) Correlation between gene-wise estimates of Tajima's D and π_N/π_S at the within-host scale. Each circle represents a gene in the H37Rv genome. π_N/π_S values are plotted on a logarithmic (base 2) scale. Blue dots are polyketide synthases that encode virulence lipids (*see Table 3*); magenta dots are toxin-antitoxin encoding genes [45]. WH – within-host; BH – between-host.

One of the signatures of positively selected genetic variants is their repeated emergence in replicate experiments during *in vitro* evolution [46], or across different patients *in vivo* [45,47]. We identified 1,500 SNPs that emerged in the within-host *M.tb* populations of all three patients with TB, and an additional 1,044 SNPs that emerged in two of the three patient *M.tb* populations (hereafter called 'convergent SNPs'). Among genes harboring SNP(s) in at least one within-host *M.tb* population, Tajima's D values (averaged across within-host samples) are lower for genes with convergent SNPs than those without (Fig. 6). This provides further support for the idea that low Tajima's D is driven by positive selection within hosts (Student's T-test, p-value = 1.1 x 10^{-109}).

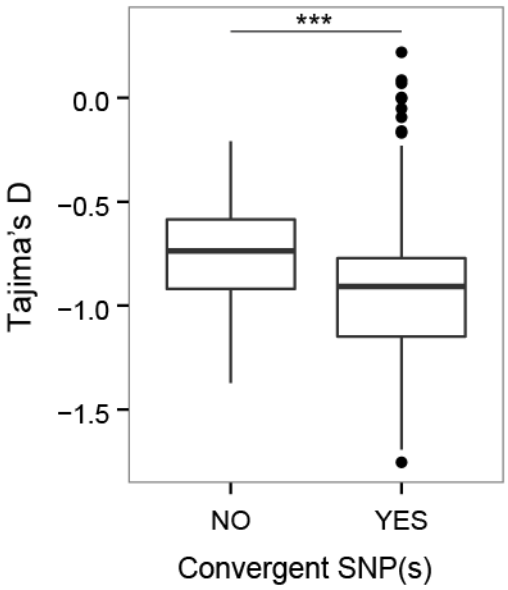


Fig 6. Tajima's D values for genes harboring convergent SNPs versus those harboring non-convergent SNPst. Average Tajima's D values (mean across within-host samples) of genes harboring one or more convergent SNPs were compared to genes harboring SNP(s) not found in multiple patient M.tb populations (*see Methods*). ***Indicates significance at the <0.001 level using a Student's t-test.

Functional categories of genes with extreme patterns of variation across evolutionary scales

Lipid metabolism genes were prominent among those with extreme patterns of variation at within- and between-host scales. Genes encoding polyketide synthases (*PKS*) are particularly striking: among 22 *PKS* homologs in the H37Rv genome (excluding the incorrectly annotated *pks15*, and counting *pks15/1* and *pks3/4* as one locus each [49]), we found 20 exhibited extreme patterns of variation (5% percentile or less, or 95% percentile or more) in one or more sample's gene-wise Tajima's D and/or π_N/π_S distribution. The most common pattern found among *PKS* is low gene-wise Tajima's D at both within- and between-host scales (Fig. 5C). In all but one of the *PKS* genes with extreme patterns of variation we also identified convergent SNPs (i.e. 19 *PKS* loci). These findings are consistent with positive selection on *PKS* during infection.

Patterns of variation in *mbtC* (*Rv2382c*) are distinct from other *PKS*. *MbtC* was not in the bottom tails of the between- or within-host, gene-wise Tajima's D distributions, nor did it harbor

any convergent SNPs. It does, however, have an extremely low π_N/π_S (1st percentile) in the between-host sample (S4 Table). This suggests that unlike the majority of PKS with evidence of positive selection within hosts, *mbtC* is under relatively strong purifying selection.

PKS-synthesized lipids are transported to the cell surface by the Mycobacterial membrane protein Large (*mmpL*) family of membrane proteins. H37Rv includes genes encoding 13 MmpL transporters [50]; substrates have not been identified for all them. MmpL have been shown to play important roles in TB pathogenesis [50,51]. *MmpL* genes were particularly notable for extreme patterns of variation: all 13 of these loci were in the extreme tail of one or more statistical measures (S4 Table). Similar to *PKS* genes, the combination of low values of gene-wise Tajima's D at the within- and between-host scales was most common. Also similar to *PKS*, 11/13 *mmpL* genes harbored convergent SNPs. Most intriguingly, *mmpL1* and *mmpL4* harbored the same SNP in a homologous region of the mmpL domain: this SNP emerged in both genes, in all three patient populations (i.e. a total of six times), and results in a non-synonymous change. Neither SNP was found in the between-host dataset.

M.tb PKS play essential roles in the biosynthesis of lipids and glycolipids of the cell envelope [49]. They are positioned at the outer edge of the envelope, at the host-pathogen interface [52]. As might be predicted based on their location, these lipids and glycolipids play important roles in the pathogenesis of TB (reviewed in [49,53,54]). PKS-synthesized lipids include mycolic acids, phthiocerol dimycocerosates (PDIM), sulfolipids (SL), polyacyl trehalose (PAT), diacyl trehalose (DAT), phenolic glycolipid (PGL), novel complex polar lipids synthesized by Pks5 (POL), and mannosyl-β-1-phosphomycoketides (MPM). They are listed in Table 3, along with the genes known to be involved in their biosynthesis and transport [49]. In addition to *PKS* and *mmpL*, several fatty acyl-AMP ligases (*FAAL*), acyl-transferases, and other

376 genes involved in these lipids' synthesis and transport exhibited extreme patterns of genetic
377 variation (**bolded** in Table 3). Patterns were more diverse than those identified among PKS (S4
378 Table).

Table 3. Polyketide synthase (PKS) synthesized virulence lipids and the genes involved in their biosynthesis and transport. Bold formatting indicates that the gene was found in the extreme of at least one samples' (between-host, a1, a2, a3, b2, c1, c2) gene-wise π_N/π_S or Tajima's D distribution.

Product	Polyketide Synthase (PKS)	fatty acyl-AMP ligase(s) (FAAL)	Acyl-transferase	Transporter	Other
phthiocerol dimycocerosates (PDIMs)	***ppsABCDE, mas***	*fadD26, fadD28*	***papA5***	***mmpL7***, *drrAB**C***	*tesA, lppX,* ***Rv2951c***, *Rv2952, Rv2953*
phenolic glycolipids (PGLs)	***pks15/1***	***fadD29, fadD22,*** *fadD28*	***papA5***	***mmpL7***	*Rv2959c, Rv2962, Rv2958c, Rv2957, tesA*
mycolic acids	***pks13***	***fadD32***			*accD4*
sulfolipids (SL)	***pks2***	*fadD23*	*papA1,* ***papA2, chp1***	***mmpL8***, *sap*	*stf0*
mannosyl-β-1-phosphomycoketides (MPM)	***pks12***				***Rv2047c,*** *Rv2049c*
polyacyltrehalose (PAT)	***pks3/4***	***fadD21***	*papA3, chp2*	***mmpL10***	
lipooligosaccharides (LOSs)	***pks5***				*Rv1500*
complex polar lipids	***pks6***	*fadD30*		***mmpL1***	

379

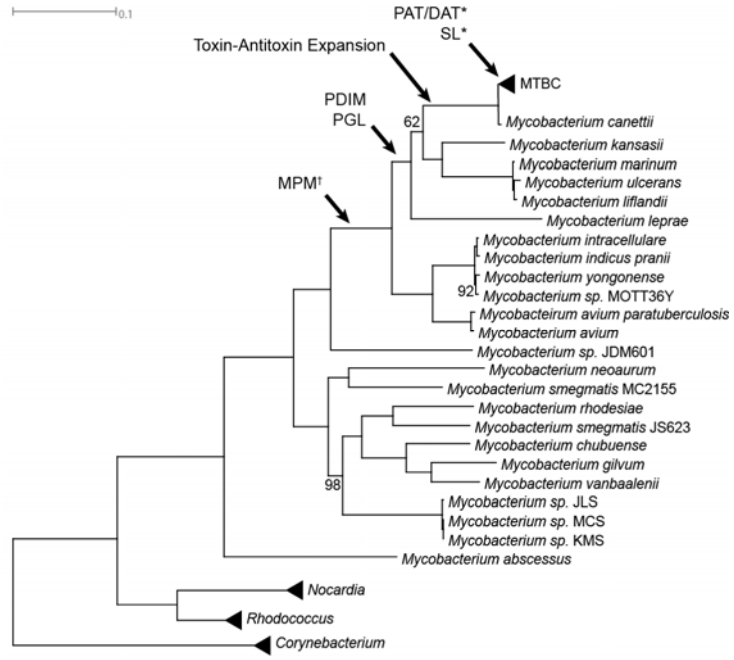
380 PDIM has major impacts on *M.tb*'s virulence in animal models of TB [55,56] via
381 numerous mechanisms (reviewed in [57]) including protection from innate immune responses
382 [58]. Synthesis of PDIM involves several PKS, i.e. Mas and PpsA-E [59]. *Mas* was in the
383 extreme low tail of the Tajima's D distributions of all within-host patient samples, and the
384 extreme high tail of the π_N/π_S distribution of two within-host samples, suggesting this gene is the
385 target of positive selection within hosts. It was also in the 7th percentile of the between-host,
386 gene-wise Tajima's D distribution (S4 Table). Interestingly, we identified identical
387 nonsynonymous SNPs (V->G) in homologous regions of *ppsC* and *mas*, within the ketoreductase

388 domain. SNPs in both genes were convergent, i.e. emerged independently across all three
389 patients so that the SNP emerged a total of six times. The convergent SNP in *mas* was found at a
390 low frequency in the between-host sample (0.5%) while the convergent SNP in *ppsC* was not
391 observed in the between-host sample. *PpsC* was in the bottom tail of the gene-wise Tajima's D
392 distribution at the between-host scale, and in four within-host samples (S4 Table). These results
393 suggest that specific loci within PKS are consistently targeted by positive selection during
394 infection. Variants under selection do not appear at the between host scale, possibly because they
395 are maladaptive for transmission, or because they remain at modest frequencies within hosts.

396 Like PDIM, the PKS-synthesized lipids PAT/DAT, SL, MPM, POL and PGL appear to
397 mediate interactions between *M.tb* and the immune system [55,60–76]. They are therefore
398 potential targets of selection to optimize *M.tb*'s immunomodulation of its host. Production of
399 PAT/DAT has been shown to vary among clinical isolates [75]; some isolates do not make them
400 at all [78]. Production of PGL also varies among clinical isolates, and variant forms of PDIM
401 have been identified in clinical isolates [54]. Our finding of extreme patterns of genetic variation
402 in multiple genes involved in their synthesis and transportation suggests PDIM, PAT/DAT, SL,
403 MPM, POL and PGL may be modified in response to environments encountered during natural
404 infection, as genetic diversity is generated within the huge population of bacteria harbored by
405 individual hosts.

406 An additional group of *M.tb* genes with distinctive patterns of variation are those
407 encoding toxin-antitoxins. Unlike the *PKS* and *mmpL* gene families, toxin-antitoxin genes are
408 characterized by high values of Tajima's D in both within- and between-host datasets, and a wide
409 range of values of π_N/π_S between hosts (Fig. 5). The *M.tb* genome includes a strikingly large
410 number of toxin-antitoxin loci, with 38-88 identified depending on the method used [45,79].

411 *M.tb*'s toxin-antitoxin systems are shared among members of the MTBC (Fig. 7), and, with one
412 exception, absent from other mycobacteria [79]. This suggests that *M.tb*'s toxin-antitoxin systems
413 serve important roles in its unique pathogenic niche; their biological roles have not, however,
414 been conclusively identified. Hypothesized functions include the modulation of growth and
415 metabolism in response to fluctuating environments and generation of 'persister' sub-populations
416 able to withstand harsh conditions [45,79,80]. Toxin-antitoxin genes have previously been shown
417 to be highly variable in their expression among clinical isolates of *M.tb*, suggesting they may be
418 responsive to fluctuating environments encountered during infection and transmission [81].



419
420 **Fig. 7. Mycobacteria maximum likelihood tree.** Phylogenetic analysis is based on a core genome
421 alignment of 57 strains (S5 Table). Arrows indicate the most probable emergence of specific lipids and
422 toxin-antitoxin expansion in the phylogenetic history of Mycobacteria based on previous studies
423 [53,58,85,87]. *Indicates lipids are only found in *M. tuberculosis*, not other species of *Mycobacterium*
424 *tuberculosis* complex (MTBC). †Indicates uncertainty in the placement because lipids of *M. sp.* JDM501
425 have not been characterized. Bootstrap values are 100 unless otherwise labeled. Scale bar indicates the
426 mean number of substitutions per site. ABREVIATIONS: PDIM – phthiocerol dimycocerosates; PGL –
427 phenolic glycolipids; SL – sulfolipids; MPM – mannosyl-β-1-phosphomycoketides; PAT – polyacyl-
428 trehalose; DAT – diacyl-trehalose
429

430 Several bacterial toxins function as mRNA interferases, which target specific cleavage
431 sites in single-stranded RNA: they are thought to regulate protein expression via degradation of

specific mRNA transcripts [83]. *M.tb* encodes an mRNA interferase belonging to the mazF family, Rv1102c, with an unusually long (i.e. specific) target. Similar to other toxin-antitoxin loci in our dataset, *rv1102c* appears to be under diversifying selection across evolutionary scales (it was in the 95th percentiles of Tajima's D both within- and between-hosts, as well as 95th percentile of between-host π_N/π_S) (Fig. 5). Zhu et al identified ten *M.tb* genes in which the Rv1102c target is under-represented, i.e. genes expressed preferentially when the toxin is active. Interestingly, one of its targets is *pks12*, a gene essential for synthesis of MPM (Table 3), thereby linking this toxin-antitoxin system to expression of an immunomodulatory lipid (Fig. 8).

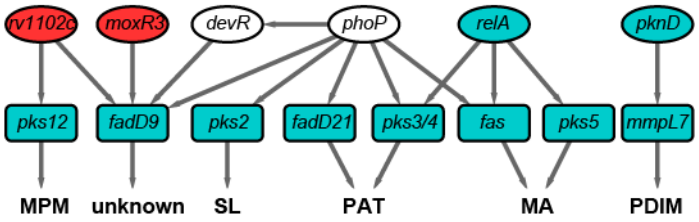


Fig. 8. Connectivity of regulators and targets involved in the synthesis of *Mycobacterium tuberculosis* immunomodulatory lipids. Circles denote genes that encode known regulators; rectangles denote genes that encode proteins involved in the synthesis or transport of the lipids at the bottom of the figure (indicated by arrows), which are targets of the regulators (indicated by arrows). Blue highlighting indicates that the gene was in the bottom 5% of the gene-wise Tajima's D distribution of the between-host and/or a within-host sample. Red highlighting indicates that the gene was in the top 5% of the gene-wise Tajima's D distribution at the between- and/or within-host scale.

Patterns of variation in *fadD9*, another putative target of Rv1102c, are very similar to *pks12* (both are in the bottom 5th percentiles of between- and within-host Tajima's D, and 95th percentile of within-host π_N/π_S) (S4 Table). Expression of *fadD9* is also under the control of two regulators that are central to *M.tb*'s adaptation to the within-host environment: PhoP [85] and DevR [84] (Fig. 8). PhoP controls the production of PAT/DAT and SL [85]. Another regulator of *fadD9*, *moxR3*, also exhibited extreme patterns of variation: this gene was in the 5th percentile of between-host πN/πS values and 95th percentile of within-host Tajima's D values (Fig. 8, S4 Table). *MoxR3* is upregulated during re-aeration of hypoxic *M.tb* cultures [86]. The function of

457 FadD9 has not been identified; patterns of variation in this gene and its regulators, as well as its
458 co-regulation with virulence associated lipids, suggest it plays an important role in host pathogen
459 interactions, and that it is a target of selection within hosts.

460 We identified several other examples of extreme patterns of genetic variation in both
461 regulators and targets involved in synthesis of immunomodulatory lipids (Fig. 8). For example, a
462 regulator of PDIM synthesis, *pknD*, was in in the low tail of between-host and within-host
463 Tajima's D (three samples). PknD is thought to regulate deposition of PDIM on the cell wall via
464 its effects on MmpL7, which transports PDIM [87,88]. Another example is the GTP
465 pyrophosphokinase *relA* (*Rv2583c*), which regulates *M.tb* gene expression during the chronic
466 phase of murine infection and plays a central role in the response to hypoxia and starvation
467 [95,95]. *RelA* was in the low tail of between-host and within-host (three samples) Tajima's D.
468 RelA modulates expression of *pks3/4*, *pks5* and *fas*, all of which had extreme values of Tajima's
469 D within- and between-hosts and harbored convergent SNPs. *Pks3/4* is essential for the
470 production of PAT [91]. *Fas* and *pks5* are in the mycolic acid superpathway. Pks5 has also been
471 implicated in synthesis of some forms of DAT [55].Taken together, these results suggest that
472 *M.tb*'s complement of immunomodulatory lipids is optimized during natural infection as a result
473 of selection for adaptive mutations in regulators, in addition to biosynthetic enzymes and
474 transporters of these lipids. It is not clear what evolutionary dynamics underlie the distinct
475 patterns of variation observed here, particularly the high Tajima's D associated with toxin-
476 antitoxin loci versus low Tajima's D in other regulators. This is an interesting topic for future
477 study.

478 Several features are conducive to efficient selection for adaptive mutations in genes
479 controlling the synthesis, regulation, and transportation of PKS synthesized lipids. There are

480 connections between biosynthetic pathways controlling production of distinct PKS synthesized
481 lipids (e.g. PDIM, SL, and mycolic acids) such that metabolites can be shuttled between them
482 [92,93]. This flexibility allows *M.tb* to respond rapidly to environmental fluctuations. It may also
483 allow more efficient selection for adaptive mutations, since single mutations can potentially
484 affect multiple lipid products. In addition, since intermediate metabolites can be shuttled down
485 multiple pathways, they need not accumulate with potentially toxic effects if one pathway is
486 affected by a harmful mutation.

487 PDIM, PAT/DAT and SL are not essential for growth of *M.tb* in artificial media, and
488 PDIM is in fact frequently lost during passage of *M.tb* in the laboratory [75,78]. This suggests
489 that functions of these lipids are specific to natural, within-host environments. The expression of
490 *M.tb* immunomodulatory lipids is responsive to physiological conditions (hypoxia, starvation)
491 encountered during natural infection [95,96]. Studies in animal models and experimental
492 macrophage infections have also shown that the shift to life within a host cell is accompanied by
493 alterations in expression of genes involved in lipid metabolism with consequent changes to the
494 cell envelope [92,97–99]. Further supporting the idea that these lipids are important for
495 adaptation to the pathogenic niche, MPM, PGL, and PDIM are only found among pathogenic
496 mycobacteria, whereas DAT/PAT and SL are specific to the *Mycobacterium tuberculosis*
497 complex (MTBC) (Fig. 7) [53,58,85]. It was shown recently that even among members of the
498 MTBC, production of DAT/PAT and SL is specific to human-pathogenic mycobacteria [85].

499 Recent work has shown that PAT/DAT, PDIM, and SL all impair phagosomal
500 acidification and thereby improve *M.tb* survival within macrophages; there appears to be
501 significant flexibility in how these functions are performed, and distinct lipid moieties may
502 compensate for each other [75]. By providing cover during exploration of the fitness landscape,

both functional redundancy and metabolic flexibility may increase the potential for rapid adaptation. For example, they would enable progress toward fitness peaks that are otherwise unreachable as a result of a preceding 'valley' created by mutations that are deleterious except in certain combinations. This may be particularly important for a clonally reproducing organism such as *M.tb*.

There are several limitations and caveats that apply to the interpretation of patterns of *M.tb* genetic variation across evolutionary scales. For example, in clonally reproducing organisms, we do not expect to observe local depressions in Tajima's D at loci under positive selection. There is little theoretical work to guide the interpretation of locus specific variation in Tajima's D for clonal organisms. We have made several observations linking positive selection during infection with low values of Tajima's D at within- and between-host scales. However, explanations other than positive selection are also possible. An alternative (though not necessarily mutually exclusive) hypothesis is that there may be a higher mutation rate among genes with low values of Tajima's D, which could also lead to an excess of rare variants. There is evidence suggesting that stress-induced mutagenesis may be directed at actively transcribed genes [100]; the genes we identified with extremely low values of Tajima's D could be transcribed preferentially during infection. Selection and genetic drift acting during cycles of transmission and infection are likely to affect patterns of *M.tb* variation in complex ways. This is an exciting area for future research, which will guide the interpretation of expanding population genomic data sets. Regardless of the underlying mechanism, the consistent identification of specific functional categories and individual genes with extreme patterns of variation across evolutionary scales and statistical measures suggests that they play important roles in survival and adaptation of *M.tb* in its natural environment.

Conclusions

Each patient with TB harbors a diverse, dynamic population of *M.tb* cells that evolves measurably over the course of their infection. Our results suggest that overall bacterial diversity is culled and re-generated as *M.tb* infections spread among hosts. In addition, despite apparent clonal reproduction, measures of diversity vary among functional groups of loci in the *M.tb* genome. We have found signatures of selection in specific genes and groups of genes that are consistent across statistical measures and evolutionary scales. Genes involved in the synthesis, transportation, and regulation of cell envelope lipids and glycolipids appear to be targets of natural selection during TB infection. These lipids are positioned at the host-pathogen interface, and are known to be critical to the immunopathogenesis of TB. Our findings suggest that they are remodeled in response to fluctuating selection pressures imposed by specific features of individual hosts, and/or changes in the bacterium's environment during cycles of transmission and infection.

Methods

Data Collection for Within- and Between-Host Datasets

Within-host *M.tb* data set (n = 7):

We used previously published WGS data [14] to characterize within-host populations of *M.tb*. Serial sputum samples from three patients with TB were collected during routine clinical care in Shanghai as previously described [14]. Seven sputum samples (three from one patient and two each from the others) were grown on Lowenstein-Jensen slants for 4 weeks, and genomic DNA from each mixed population sample was sequenced on the Illumina platform (pool-seq).

549 Sequence data were submitted to the National Center for Biotechnology Information Sequence
550 Read Archive (NCBI SRA) under Accession No. SRA050212. Since sequencing error biases are
551 known to vary across platforms [101], we only used WGS data generated on the Illumina Hi-Seq
552 (data from a variety of modalities were analyzed in the original Sun *et al.* (2012) study).

553 Between-host *M.tb* data set (n = 201):

554 We used previously published WGS data from 201 diverse, globally extant strains of
555 *M.tb* [29] to characterize between-host populations of *M.tb*. The data set includes *M.tb* isolates
556 from all seven major lineages of *M.tb* [102,103]. Accession numbers and more detailed
557 information about the strains are in S2 Table.

558 *Processing of raw sequencing reads*

559 For the within-host *M.tb* data set, we trimmed low-quality bases from the FASTQ data
560 using a threshold quality score of 20, and reads of length less than 40bp were discarded using
561 PoPoolation software [18]. We mapped reads to H37Rv [30] using the default parameters of the
562 BWA MEM algorithm with the –M flag enabled for downstream compatibility [104], and we
563 removed duplicates using Picard Tools (http://picard.sourceforge.net). Local realignment was
564 performed with the Genome Analysis ToolKit (GATK) [105], and aligned reads were converted
565 to mpileup format using Samtools software [106]. The resulting reference-guided assembly of
566 each sample spanned over 97% of the H37Rv genome, with a mean depth of coverage per site
567 ranging from 38X to 312X (Table 1). Loci at which indels were present in at least three reads of
568 a sample were removed along with 5bp of flanking sequence using the PoPoolation package
569 [18].

570 For the between-host *M.tb* data set, we trimmed low-quality bases from FASTQ data
571 using a threshold quality of 15, and reads resulting in less than 20bp length were discarded using

572 Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) - a wrapper tool
573 around Cutadapt [105] and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
574 Reads were mapped to H37Rv [30] using the suggested algorithm of BWA for the given
575 sequencing strategy (e.g. paired-end/single-end, read length) [104,108], and duplicates were
576 removed using Picard Tools (http://picard.sourceforge.net). We used the GATK to perform local
577 realignment and variant calls using a minimum phred-scaled confidence threshold of 20 [105].
578 Genome alignments were generated with in-house scripts and can be found at
579 https://github.com/tracysmith/RGAPepPipe.

580 Transposable elements, phage elements, and repetitive families of genes (PE, PPE, and
581 PE-PGRS gene families) that are poorly resolved with short read sequencing were removed from
582 the mpileup files and alignment prior to subsequent analyses plus 100bp up- and down-stream of
583 the genes. We additionally removed regions that were found to have poor mapping quality using
584 the CallableLoci tool of the GATK: for each within-host sample, any region reported as poorly
585 mapped using the following flags was removed from all datasets (including between-host) plus
586 5bp up- and down-stream: -frlmq 0.04 –mmq 20 -mlmq 19. As an additional measure, we also
587 removed sites with the lowest 2% average mapping quality in the global alignment from all
588 datasets prior to subsequent analysis.

589

590 *Population Genetic Estimates of Within-host Populations*

591 We used the PoPoolation package [18] to estimate nucleotide diversity (π), Watterson's
592 theta (θ_W), and Tajima's D in sliding windows across the *M.tb* genome. Sensitivity analyses of
593 these statistics to input parameters are described in the S7 File. Parameter estimates were
594 strongly influenced by sequencing coverage (S7 File). In order to alleviate biases in parameter

595 estimates caused by variable coverage among samples (Table 1), we randomly sub-sampled read
596 data without replacement to a uniform coverage of 50X; this process was performed 10 times for
597 each sample to reduce potential biases introduced by sampling of rare alleles. We used default
598 equations in the PoPoolation package to estimate π, θ_W, and Tajima's D; these estimators apply
599 corrections for sequencing errors, which are difficult to distinguish from rare alleles in pooled re-
600 sequencing data [18,19]. Unless otherwise noted, all calculations performed with the PoPoolation
601 package were implemented using the recommended minimum minor allele count of 2 for 50X
602 coverage, and a pool-size of 10,000 (*see S7 File for justification*). Genome-wide estimates for
603 each sample are reported as the mean across the 10 replicate sub-sampled mpileups. For
604 visualization, sliding-windows of 100K with a step-size of 10K were used (Fig. 1). Because of
605 the low sequence coverage obtained for sample b1 (Table 1), we estimated sliding-window
606 values of π, θ_W, and Tajima's D based on sub-sampling to a uniform coverage of 25X for both
607 samples obtained from patient B (b1 and b2), allowing us to compare these two time points (S1
608 Figure).

609 To calculate gene-by-gene estimates of π, θ, and Tajima's D from the 50X sub-sampled
610 mpileups, we again used the PoPoolation package [18]. Gene coordinates were obtained from the
611 standardized Gene Transfer Format (.gtf) for the H37Rv annotation on the TB database
612 (tbdb.org) [109]. Excluding genes with inadequate coverage (<55% of the gene), we calculated
613 the mean value of each statistic across the 10 replicate sub-samplings for each gene of each
614 sample, and compared it to all other genes within the sample.

615 Using the same .gtf file and the PoPoolation package, we also calculated the average
616 number of nonsynonymous differences per nonsynonymous site (π_N) and the average number of
617 synonymous differences per synonymous site (π_S) for each gene in the 50X sub-sampled

mpileups. Recognizing that chance samplings of very rare mutations in one replicate sub-sampled mpileup would lead to skewed distributions, we took the median values of π_N and π_S across 9 replicate sub-samplings for each gene of each sample and calculated π_N/π_S. Excluding genes with inadequate coverage (<55% of the gene) and genes with π_N and/or π_S equal to zero, π_N/π_S values were compared relative to all genes passing these criteria within the sample.

Treating each temporal sample of a patient as a population, we used PoPoolation2 [110] to estimate F_{ST} for each variable site in the genome. To reduce biases resulting from variable coverage, we merged randomly sub-sampled pileup files and calculated F_{ST} and allele frequency changes for each polymorphic site containing at least 20 counts of the minor allele (patients A and C), and at least 10 counts for comparisons among longitudinal populations of patient B. This is consistent with a minimum minor allele count of 2 for 50X coverage scaled accordingly for the 10 replicate sub-samplings of 50X for *M.tb* population samples from patients A and C and 25X for *M.tb* population samples from patient B that were merged prior to analysis. We made only one comparison for patients B and C, as there are only two temporal samples; for Patient A, we calculated F_{ST} for each pairwise population comparison (i.e. a1-a2, a1-a3, a2-a3). SNPs falling in the top 0.1% of each pairwise comparison are annotated in S5 Table.

Population Genetics of Globally Extant Strains (Between-Host)

We used EggLib [21] to estimate π and θ_W from whole genome alignments of all 201 globally extant *M.tb* strains, *M.tb* isolates from lineage 2 (East Asian lineage, n = 37), *M.tb* isolates from Chinese-born TB patients in the global survey (n = 14), and *M.tb* isolates from individuals born in East Asia (n = 18) (S2 Table) [29]. We required a minimum of 75% of the strains have non-missing data for a site to be included in the analysis.

We calculated gene-by-gene estimates of π, □, Tajima's D, π_N, and π_S for the global dataset in the same way as was done for within-host samples, save for a few exceptions: rather than use the default PoPoolation estimators that apply corrections for sequencing errors [18,19], we employed the "disable-corrections" flag (calculations are performed using the classical equations), set the "minimum minor allele count" to one, and the "pool-size" was set to 201 to reflect the number of strains in the dataset; the "minimum coverage" was set to 151 to exclude any genes that were not represented by at least 75% of strains in the dataset. Finally, no averaging was performed. For π_N/π_S calculations, we again excluded genes with π_N and/or π_S equal to zero, and compared values relative to all genes passing these criteria within the sample.

Functional & Pathway Enrichment Analyses

For each within-host sample, between 81.8-91.8% of annotated genes in the H37Rv genome had sufficient coverage to calculate Tajima's D; for the between-host sample, Tajima's D was calculated for 91.7% of annotated genes in the H37Rv genome, as this was the fraction covered by at least 75% of the strains. Genes with Tajima's D values in the top and bottom 5% of the distribution for a given sample were deemed candidate genes of selection. (Note that sample b1 was not included in this analysis due to low coverage.) The significance of enrichment for functional categories in candidate genes of selection was assessed with a two-sided Fisher's exact test. To account for multiple hypothesis testing, we used a false discovery rate of 5% and calculated *q*-values (Stats Package, [111]). We used the following annotation categories to classify *M.tb* genes: computationally predicted Clusters of Orthologous Groups (COG) (n = 21 categories) [28]; essential and nonessential genes for growth *in vitro* as determined by transposon site hybridization (TraSH) mutagenesis [112]; genes essential for growth in a murine

model of TB [31]; "targets of independent mutation" associated with drug resistance [35]; and the *M.tb*-specific, manually curated functional annotation lists from TubercuList (n = 7) [27]. COG annotations for the H37Rv genome were obtained from the TB database (tbdb.org) [109], as were the TraSH "*in vitro* essential", "*in vivo* essential", and "nonessential" gene annotations. TubercuList functional annotations were obtained from tuberculist.epfl.ch [27] and reflect the most up-to-date annotations when the database was accessed (12/01/2013). We did not include the TubercuList categories "PE/PPE", "stable RNA", and "unknown" in our analyses. *Q*-values used to generate Fig. 2 are reported in S2 Table.

Only 3.5-28% of annotated genes in the H37Rv genome had sufficient coverage and contained both nonsynonymous and synonymous variation in a given within-host sample (not including sample b1). We reasoned that genes in the lower tail of the π_N/π_S distribution of each within-host sample are likely not to be representative of genes under the strongest purifying selection, as any gene containing no nonsynonymous variation would result in a π_N/π_S value of zero regardless of whether it contained high or low synonymous diversity. While a large percentage of genes were also excluded from the analysis because they lacked synonymous variation, the inference to be drawn from an elevated π_N/π_S value still holds. Thus, genes with π_N/π_S values in the top 1% of the distribution for a given sample were classified as outliers. To look for commonalities among patients, we also examined all genes found to have a positive π_N/π_S value in at least one sample from each patient. Upon noticing a preponderance of genes in the superpathway of mycolate biosynthesis on the cellular overview tool of the TB database (tbdb.org) [109] we formally tested for an enrichment of genes in this category using a two-sided Fisher's exact test.

687 *Correlations among statistics*

688 To look at the correlations among statistics across evolutionary scales, we summarized
689 the statistics estimated for the within-host samples. For Tajima's D, the mean across all seven
690 within-host samples was used; for $\pi_N/\pi S$ the nonsynonymous and synonymous variation were
691 summed across all within-host samples and then divided. R (Stats Package, [111]) was used to
692 calculate Pearson's correlation coefficients and *p*-values for each comparison.

693

694 *Convergent SNPs*

695 To identify SNPs that emerged in multiple patient *M.tb* populations, only sites covered at
696 ≥ 50X in at least one sample from the *M.tb* populations of patient A and C, ≥ 25X in at least one
697 sample from the *M.tb* population of patient B, and in ≥ 75% of global isolates in the between-
698 host dataset were considered, resulting in 3,834,551bp of the H37Rv genome being included in
699 the analysis. SNPs had to be found at a minimum frequency of 0.04 in the sub-sampled data
700 (merged across replicate sub-samplings – same data as was used for F_{ST} outlier analysis) of the
701 within-host samples to be considered a SNP. 7,801 unique SNPs were identified under these
702 criteria across all within-host samples. SNPs were deemed "convergent" if the same nucleotide
703 change occurred in at least one sample from 2 or 3 different patient *M.tb* populations. (Two
704 SNPs were excluded from the analysis based on a different nucleotide change occurring in
705 different populations.) This resulted in the identification of 1500 SNPs found in at least one
706 sample from all 3 patients' *M.tb* populations, and an additional 1,044 SNPs found in at least one
707 sample from 2 of the 3 patients' *M.tb* populations. A two-sided Student's t-test was used to
708 determine if, on average, genes harboring a convergent SNP had statistically different Tajima's
709 D values than genes harboring at least one SNP in a within-host sample, but no convergent SNP.

Mycobacteria Core Genome Alignment and Phylogenetic Tree

We downloaded finished genomes of mycobacteria species from NCBI (S5 Table). Reference guided assemblies for unfinished genomes were performed as described above for the between-host dataset. We used Prokka v 1.7 [113] for genome annotation. Protein sequences output by Prokka were clustered into orthologous groups using OrthoMCL [114]. The core proteins (those found only once in every genome) were aligned using MAFFT [115], trimmed with trimAl [116], and concatenated. Scripts used for core genome analysis can be found at https://github.com/tatumdmortimer/core-genome-alignment. We used RAxML v 8 [115] for maximum likelihood phylogenetic analysis of the core alignment. We used Dendroscope [118] for tree viewing and editing.

Authors' contributions

CSP and MBO designed the analyses, which were done by MBO save for the Mycobacteria species tree which was made by TDM. CSP conceived of the study. MBO and CSP wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Jennifer Bratburd for her contribution to the generation of the Mycobacteria species tree, and Tracy M. Smith for her input on the manuscript. This work was supported by start-up funds provided to CSP by the University of Wisconsin-Madison and the University of Wisconsin Medical Foundation. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program (DGE-1255259) to MBO and TDM,

733 and the National Institutes of Health sponsored University of Wisconsin Training Grant in
734 Genetics (T32 GM07133) to MBO.

References

1. WHO | Global tuberculosis report 2014 (n.d.). WHO. Available: http://www.who.int/tb/publications/global_report/en/. Accessed 6 January 2015.

2. Klopper M, Warren RM, Hayes C, Gey van Pittius NC, Streicher EM, et al. (2013) Emergence and Spread of Extensively and Totally Drug-Resistant Tuberculosis, South Africa. Emerg Infect Dis 19: 449–455. doi:10.3201/eid1903.120246.

3. Achtman M (2012) Insights from genomic comparisons of genetically monomorphic bacterial pathogens. Philos Trans R Soc B Biol Sci 367: 860–867. doi:10.1098/rstb.2011.0303.

4. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, et al. (2013) The Role of Selection in Shaping Diversity of Natural M. tuberculosis Populations. PLoS Pathog 9: e1003543. doi:10.1371/journal.ppat.1003543.

5. Pepperell C, Hoeppner VH, Lipatov M, Wobeser W, Schoolnik GK, et al. (2010) Bacterial Genetic Signatures of Human Social Phenomena among M. tuberculosis from an Aboriginal Canadian Population. Mol Biol Evol 27: 427–440. doi:10.1093/molbev/msp261.

6. Pepperell CS, Granka JM, Alexander DC, Behr MA, Chui L, et al. (2011) Dispersal of Mycobacterium tuberculosis via the Canadian fur trade. Proc Natl Acad Sci 108: 6526–6531. doi:10.1073/pnas.1016708108.

7. Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EPC (2012) After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. Genome Res 22: 721–734. doi:10.1101/gr.129544.111.

8. Tanaka MM (2004) Evidence for positive selection on Mycobacterium tuberculosis within patients. BMC Evol Biol 4: 31. doi:10.1186/1471-2148-4-31.

9. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, et al. (2013) Within-Host Evolution of Staphylococcus aureus during Asymptomatic Carriage. PLoS ONE 8: e61319. doi:10.1371/journal.pone.0061319.

10. McAdam PR, Holmes A, Templeton KE, Fitzgerald JR (2011) Adaptive Evolution of Staphylococcus aureus during Chronic Endobronchial Infection of a Cystic Fibrosis Patient. PLoS ONE 6. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3166311/. Accessed 1 August 2013.

11. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, et al. (2011) Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci 108: 5033–5038. doi:10.1073/pnas.1018444108.

12. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, et al. (2011) Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. Nat Genet 43: 482–486. doi:10.1038/ng.811.

13. Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, et al. (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet 43: 1275–1280. doi:10.1038/ng.997.

14. Sun G, Luo T, Yang C, Dong X, Li J, et al. (2012) Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients. J Infect Dis 206: 1724–1733. doi:10.1093/infdis/jis601.

15. Saunders NJ, Trivedi UH, Thomson ML, Doig C, Laurenson IF, et al. (2011) Deep resequencing of serial sputum isolates of Mycobacterium tuberculosis during therapeutic failure due to poor compliance reveals stepwise mutation of key resistance genes on an otherwise stable genetic background. J Infect 62: 212–217. doi:10.1016/j.jinf.2011.01.003.

16. Eldholm V, Norheim G, von der Lippe B, Kinander W, Dahle UR, et al. (2014) Evolution of extensively drug-resistant Mycobacterium tuberculosis from a susceptible ancestor in a single patient. Genome Biol 15: 490.

17. Merker M, Kohl TA, Roetzer A, Truebe L, Richter E, et al. (2013) Whole Genome Sequencing Reveals Complex Evolution Patterns of Multidrug-Resistant Mycobacterium tuberculosis Beijing Strains in Patients. PLoS ONE 8: e82551. doi:10.1371/journal.pone.0082551.

18. Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, et al. (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PloS One 6: e15925. doi:10.1371/journal.pone.0015925.

19. Achaz G (2008) Testing for Neutrality in Samples With Sequencing Errors. Genetics 179: 1409–1424. doi:10.1534/genetics.107.082198.

20. Supply P, Warren RM, Bañuls A-L, Lesjean S, Van Der Spuy GD, et al. (2003) Linkage disequilibrium between minisatellite loci supports clonal evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area. Mol Microbiol 47: 529–538. doi:10.1046/j.1365-2958.2003.03315.x.

21. Mita SD, Siol M (2012) EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet 13: 27. doi:10.1186/1471-2156-13-27.

22. Cohen T, Helden PD van, Wilson D, Colijn C, McLaughlin MM, et al. (2012) Mixed-Strain Mycobacterium tuberculosis Infections and the Implications for Tuberculosis Treatment and Control. Clin Microbiol Rev 25: 708–719. doi:10.1128/CMR.00021-12.

23. Iseman MD (2000) A clinician's guide to tuberculosis. 1st ed. Philadelphia: Lippincott Williams & Wilkins. 460 p.

24. Bloom B (1994) Tuberculosis: Pathogenesis, Protection and Control. 1st ed. Washington, DC: American Society for Microbiology. 653 p.

25. Namouchi A, Karboul A, Fabre M, Gutierrez MC, Mardassi H (2013) Evolution of Smooth Tubercle Bacilli PE and PE_PGRS Genes: Evidence for a Prominent Role of Recombination and Imprint of Positive Selection. PLoS ONE 8: e64718. doi:10.1371/journal.pone.0064718.

26. Tajima F (1989) Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123: 585–595.

27. Lew JM, Kapopoulou A, Jones LM, Cole ST (2011) TubercuList – 10 years after. Tuberculosis 91: 1–7. doi:10.1016/j.tube.2010.09.008.

28. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28: 33–36. doi:10.1093/nar/28.1.33.

29. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, et al. (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet advance online publication. Available: http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.2744.html. Accessed 3 September 2013.

30. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, et al. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393: 537–544. doi:10.1038/31159.

31. Sassetti CM, Rubin EJ (2003) Genetic requirements for mycobacterial survival during infection. Proc Natl Acad Sci 100: 12989–12994. doi:10.1073/pnas.2134250100.

32. Shimono N, Morici L, Casali N, Cantrell S, Sidders B, et al. (2003) Hypervirulent mutant of Mycobacterium tuberculosis resulting from disruption of the mce1 operon. Proc Natl Acad Sci 100: 15918–15923. doi:10.1073/pnas.2433882100.

33. Stavrum R, Stavrum A-K, Valvatne H, Riley LW, Ulvestad E, et al. (2011) Modulation of Transcriptional and Inflammatory Responses in Murine Macrophages by the Mycobacterium tuberculosis Mammalian Cell Entry (Mce) 1 Complex. PLoS ONE 6: e26295. doi:10.1371/journal.pone.0026295.

34. Warner DF, Mizrahi V (2013) Complex genetics of drug resistance in Mycobacterium tuberculosis. Nat Genet 45: 1107–1108. doi:10.1038/ng.2769.

35. Silva PEAD, Palomino JC (2011) Molecular basis and mechanisms of drug resistance in Mycobacterium tuberculosis: classical and new drugs. J Antimicrob Chemother 66: 1417–1430. doi:10.1093/jac/dkr173.

36. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, et al. (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet 45: 1183–1189. doi:10.1038/ng.2747.

37. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340–345. doi:10.1038/ng.78.

38. Myles S, Tang K, Somel M, Green RE, Kelso J, et al. (2008) Identification and Analysis of Genomic Regions with Large Between-Population Differentiation in Humans. Ann Hum Genet 72: 99–110. doi:10.1111/j.1469-1809.2007.00390.x.

39. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320. doi:10.1038/nature04226.

40. Gavalda S, Léger M, Rest B van der, Stella A, Bardou F, et al. (2009) The Pks13/FadD32 Crosstalk for the Biosynthesis of Mycolic Acids in Mycobacterium tuberculosis. J Biol Chem 284: 19255–19264. doi:10.1074/jbc.M109.006940.

41. Kawate T, Iwase N, Shimizu M, Stanley SA, Wellington S, et al. (2013) Synthesis and structure–activity relationships of phenyl-substituted coumarins with anti-tubercular activity that target FadD32. Bioorg Med Chem Lett 23: 6052–6059. doi:10.1016/j.bmcl.2013.09.035.

42. Stanley SA, Kawate T, Iwase N, Shimizu M, Clatworthy AE, et al. (2013) Diarylcoumarins inhibit mycolic acid biosynthesis and kill Mycobacterium tuberculosis by targeting FadD32. Proc Natl Acad Sci 110: 11565–11570. doi:10.1073/pnas.1302114110.

43. Chalut C, Botella L, Sousa-D'Auria C de, Houssin C, Guilhot C (2006) The nonredundant roles of two 4′-phosphopantetheinyl transferases in vital processes of Mycobacteria. Proc Natl Acad Sci 103: 8511–8516. doi:10.1073/pnas.0511129103.

44. Quadri LEN, Sello J, Keating TA, Weinreb PH, Walsh CT (1998) Identification of a Mycobacterium tuberculosis gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. Chem Biol 5: 631–645. doi:10.1016/S1074-5521(98)90291-5.

45. Gupta A (2009) Killing activity and rescue function of genome-wide toxin–antitoxin loci of Mycobacterium tuberculosis. FEMS Microbiol Lett 290: 45–53. doi:10.1111/j.1574-6968.2008.01400.x.

46. Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet 4: 457–469. doi:10.1038/nrg1088.

47. Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, et al. (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet 43: 1275–1280. doi:10.1038/ng.997.

48. Marvig RL, Sommer LM, Molin S, Johansen HK (2014) Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat Genet. Available: http://www.nature.com/doifinder/10.1038/ng.3148. Accessed 22 November 2014.

49. Quadri LEN (2014) Biosynthesis of mycobacterial lipids by polyketide synthases and beyond. Crit Rev Biochem Mol Biol 49: 179–211. doi:10.3109/10409238.2014.896859.

50. Domenech P, Reed MB, Barry CE (2005) Contribution of the Mycobacterium tuberculosis MmpL protein family to virulence and drug resistance. Infect Immun 73: 3492–3501. doi:10.1128/IAI.73.6.3492-3501.2005.

51. MacGurn JA, Cox JS (2007) A Genetic Screen for Mycobacterium tuberculosis Mutants Defective for Phagosome Maturation Arrest Identifies Components of the ESX-1 Secretion System. Infect Immun 75: 2668–2678. doi:10.1128/IAI.01872-06.

52. Domenech P, Reed MB, Dowd CS, Manca C, Kaplan G, et al. (2004) The Role of MmpL8 in Sulfatide Biogenesis and Virulence of Mycobacterium tuberculosis. J Biol Chem 279: 21257–21265. doi:10.1074/jbc.M400324200.

53. Jackson M, Stadthagen G, Gicquel B (2007) Long-chain multiple methyl-branched fatty acid-containing lipids of Mycobacterium tuberculosis: Biosynthesis, transport, regulation and biological activities. Tuberculosis 87: 78–86. doi:10.1016/j.tube.2006.05.003.

54. Neyrolles O, Guilhot C (2011) Recent advances in deciphering the contribution of Mycobacterium tuberculosis lipids to pathogenesis. Tuberculosis 91: 187–195. doi:10.1016/j.tube.2011.01.002.

55. Camacho LR, Ensergueix D, Perez E, Gicquel B, Guilhot C (1999) Identification of a virulence gene cluster of Mycobacterium tuberculosis by signature-tagged transposon mutagenesis. Mol Microbiol 34: 257–267. doi:10.1046/j.1365-2958.1999.01593.x.

56. Cox JS, Chen B, McNeil M, Jacobs WR (1999) Complex lipid determines tissue-specific replication of Mycobacterium tuberculosis in mice. Nature 402: 79–83. doi:10.1038/47042.

57. Kaufmann SHE, Rubin (2008) Handbook of tuberculosis: molecular biology and biochemistry. Weinheim: Wiley-VCH.

58. Day TA, Mittler JE, Nixon MR, Thompson C, Miner MD, et al. (2014) Mycobacterium tuberculosis strains lacking the surface lipid phthiocerol dimycocerosate are susceptible to killing by an early innate host response. Infect Immun: IAI.01340–13. doi:10.1128/IAI.01340-13.

59. Mathur M, Kolattukudy PE (1992) Molecular cloning and sequencing of the gene for mycocerosic acid synthase, a novel fatty acid elongating multifunctional enzyme, from Mycobacterium tuberculosis var. bovis Bacillus Calmette-Guerin. J Biol Chem 267: 19388–19395.

60. Lemassu A, Lanéelle M-A, Daffé M (1991) Revised structure of a trehalose-containing immunoreactive glycolipid of Mycobacterium tuberculosis. FEMS Microbiol Lett 78: 171–176. doi:10.1111/j.1574-6968.1991.tb04438.x.

61. Minnikin DE, Dobson G, Sesardic D, Ridell M (1985) Mycolipenates and Mycolipanolates of Trehalose from Mycobacterium tuberculosis. J Gen Microbiol 131: 1369–1374. doi:10.1099/00221287-131-6-1369.

62. Rousseau C, Neyrolles O, Bordat Y, Giroux S, Sirakova TD, et al. (2003) Deficiency in mycolipenate- and mycosanoate-derived acyltrehaloses enhances early interactions of Mycobacterium tuberculosis with host cells. Cell Microbiol 5: 405–415. doi:10.1046/j.1462-5822.2003.00289.x.

63. Saavedra R, Segura E, Leyva R, Esparza LA, López-Marín LM (2001) Mycobacterial Di-O-Acyl-Trehalose Inhibits Mitogen- and Antigen-Induced Proliferation of Murine T Cells In Vitro. Clin Diagn Lab Immunol 8: 1081–1088. doi:10.1128/CDLI.8.6.1-91-1088.2001.

64. Lee K-S, Dubey VS, Kolattukudy PE, Song C-H, Shin A-R, et al. (2007) Diacyltrehalose of Mycobacterium tuberculosis inhibits lipopolysaccharide- and mycobacteria-induced proinflammatory cytokine production in human monocytic cells. FEMS Microbiol Lett 267: 121–128. doi:10.1111/j.1574-6968.2006.00553.x.

65. Goren MB, D'Arcy Hart P, Young MR, Armstrong JA (1976) Prevention of phagosome-lysosome fusion in cultured macrophages by sulfatides of Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 73: 2510–2514.

66. Brozna JP, Horan M, Rademacher JM, Pabst KM, Pabst MJ (1991) Monocyte responses to sulfatide from Mycobacterium tuberculosis: inhibition of priming for enhanced release of superoxide, associated with increased secretion of interleukin-1 and tumor necrosis factor alpha, and altered protein phosphorylation. Infect Immun 59: 2542–2548.

67. Brodin P, Poquet Y, Levillain F, Peguillet I, Larrouy-Maumus G, et al. (2010) High Content Phenotypic Cell-Based Visual Screen Identifies Mycobacterium tuberculosis Acyltrehalose-Containing Glycolipids Involved in Phagosome Remodeling. PLoS Pathog 6: e1001100. doi:10.1371/journal.ppat.1001100.

68. Moody DB, Ulrichs T, Mühlecker W, Young DC, Gurcha SS, et al. (2000) CD1c-mediated T-cell recognition of isoprenoid glycolipids in Mycobacterium tuberculosis infection. Nature 404: 884–888. doi:10.1038/35009119.

69. Matsunaga I, Bhatt A, Young DC, Cheng T-Y, Eyles SJ, et al. (2004) Mycobacterium tuberculosis pks12 Produces a Novel Polyketide Presented by CD1c to T Cells. J Exp Med 200: 1559–1569. doi:10.1084/jem.20041429.

70. Sirakova TD, Dubey VS, Kim H-J, Cynamon MH, Kolattukudy PE (2003) The Largest Open Reading Frame (pks12) in the Mycobacterium tuberculosis Genome Is Involved in Pathogenesis and Dimycocerosyl Phthiocerol Synthesis. Infect Immun 71: 3794–3801. doi:10.1128/IAI.71.7.3794-3801.2003.

71. Ly D, Kasmar AG, Cheng T-Y, Jong A de, Huang S, et al. (2013) CD1c tetramers detect ex vivo T cell responses to processed phosphomycoketide antigens. J Exp Med 210: 729–741. doi:10.1084/jem.20120624.

72. Lynett J, Stokes RW (2007) Selection of transposon mutants of Mycobacterium tuberculosis with increased macrophage infectivity identifies fadD23 to be involved in sulfolipid production and association with macrophages. Microbiology 153: 3133–3140. doi:10.1099/mic.0.2007/007864-0.

73. Rosas-Magallanes V, Stadthagen-Gomez G, Rauzier J, Barreiro LB, Tailleux L, et al. (2007) Signature-Tagged Transposon Mutagenesis Identifies Novel Mycobacterium tuberculosis Genes Involved in the Parasitism of Human Macrophages. Infect Immun 75: 504–507. doi:10.1128/IAI.00058-06.

74. Beaulieu AM, Rath P, Imhof M, Siddall ME, Roberts J, et al. (2010) Genome-Wide Screen for Mycobacterium tuberculosis Genes That Regulate Host Immunity. PLoS ONE 5. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3000826/. Accessed 11 October 2014.

75. Dutta NK, Mehra S, Didier PJ, Roy CJ, Doyle LA, et al. (2010) Genetic Requirements for the Survival of Tubercle Bacilli in Primates. J Infect Dis 201: 1743–1752. doi:10.1086/652497.

76. Reed MB, Domenech P, Manca C, Su H, Barczak AK, et al. (2004) A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. Nature 431: 84–87. doi:10.1038/nature02837.

77. Passemar C, Arbués A, Malaga W, Mercier I, Moreau F, et al. (2014) Multiple deletions in the polyketide synthase gene repertoire of Mycobacterium tuberculosis reveal functional overlap of cell envelope lipids in host–pathogen interactions. Cell Microbiol 16: 195–213. doi:10.1111/cmi.12214.

78. Asensio JG, Maia C, Ferrer NL, Barilone N, Laval F, et al. (2006) The Virulence-associated Two-component PhoP-PhoR System Controls the Biosynthesis of Polyketide-derived Lipids in Mycobacterium tuberculosis. J Biol Chem 281: 1313–1316. doi:10.1074/jbc.C500388200.

79. Ramage HR, Connolly LE, Cox JS (2009) Comprehensive Functional Analysis of Mycobacterium tuberculosis Toxin-Antitoxin Systems: Implications for Pathogenesis, Stress Responses, and Evolution. PLoS Genet 5. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2781298/. Accessed 12 August 2014.

80. Arcus VL, Rainey PB, Turner SJ (2005) The PIN-domain toxin–antitoxin array in mycobacteria. Trends Microbiol 13: 360–365. doi:10.1016/j.tim.2005.06.008.

81. Rose G, Cortes T, Comas I, Coscolla M, Gagneux S, et al. (2013) Mapping of Genotype–Phenotype Diversity among Clinical Isolates of Mycobacterium tuberculosis by Sequence-Based Transcriptional Profiling. Genome Biol Evol 5: 1849–1862. doi:10.1093/gbe/evt138.

82. Gonzalo-Asensio J, Mostowy S, Harders-Westerveen J, Huygen K, Hernández-Pando R, et al. (2008) PhoP: A Missing Piece in the Intricate Puzzle of Mycobacterium tuberculosis Virulence. PLoS ONE 3: e3496. doi:10.1371/journal.pone.0003496.

83. Zhu L, Phadtare S, Nariya H, Ouyang M, Husson RN, et al. (2008) The mRNA interferases, MazF-mt3 and MazF-mt7 from *Mycobacterium tuberculosis* target unique pentad sequences in single-stranded RNA. Mol Microbiol 69: 559–569. doi:10.1111/j.1365-2958.2008.06284.x.

84. Guo M, Feng H, Zhang J, Wang W, Wang Y, et al. (2009) Dissecting transcription regulatory pathways through a new bacterial one-hybrid reporter system. Genome Res 19: 1301–1308. doi:10.1101/gr.086595.108.

85. Gonzalo-Asensio J, Malaga W, Pawlik A, Astarie-Dequeker C, Passemar C, et al. (2014) Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. Proc Natl Acad Sci: 201406693. doi:10.1073/pnas.1406693111.

86. Sherrid AM, Rustad TR, Cangelosi GA, Sherman DR (2010) Characterization of a Clp Protease Gene Regulator and the Reaeration Response in Mycobacterium tuberculosis. PLoS ONE 5: e11622. doi:10.1371/journal.pone.0011622.

87. Pérez J, Garcia R, Bach H, de Waard JH, Jacobs Jr. WR, et al. (2006) Mycobacterium tuberculosis transporter MmpL7 is a potential substrate for kinase PknD. Biochem Biophys Res Commun 348: 6–12. doi:10.1016/j.bbrc.2006.06.164.

88. Gomez-Velasco A, Bach H, Rana AK, Cox LR, Bhatt A, et al. (2013) Disruption of the serine/threonine protein kinase H affects phthiocerol dimycocerosates synthesis in Mycobacterium tuberculosis. Microbiology 159: 726–736. doi:10.1099/mic.0.062067-0.

89. Primm TP, Andersen SJ, Mizrahi V, Avarbock D, Rubin H, et al. (2000) The Stringent Response of Mycobacterium tuberculosis Is Required for Long-Term Survival. J Bacteriol 182: 4889–4898. doi:10.1128/JB.182.17.4889-4898.2000.

90. Dahl JL, Kraus CN, Boshoff HIM, Doan B, Foley K, et al. (2003) The role of RelMtb-mediated adaptation to stationary phase in long-term persistence of Mycobacterium tuberculosis in mice. Proc Natl Acad Sci 100: 10026–10031. doi:10.1073/pnas.1631248100.

91. Dubey VS, Sirakova TD, Kolattukudy PE (2002) Disruption of msl3 abolishes the synthesis of mycolipanoic and mycolipenic acids required for polyacyltrehalose synthesis in Mycobacterium tuberculosis H37Rv and causes cell aggregation. Mol Microbiol 45: 1451–1459. doi:10.1046/j.1365-2958.2002.03119.x.

92. Jain M, Petzold CJ, Schelle MW, Leavell MD, Mougous JD, et al. (2007) Lipidomics reveals control of Mycobacterium tuberculosis virulence lipids via metabolic coupling. Proc Natl Acad Sci 104: 5133–5138. doi:10.1073/pnas.0610634104.

93. Kruh NA, Borgaro JG, Ruzsicska BP, Xu H, Tonge PJ (2008) A Novel Interaction Linking the FAS-II and Phthiocerol Dimycocerosate (PDIM) Biosynthetic Pathways. J Biol Chem 283: 31719–31725. doi:10.1074/jbc.M802169200.

94. Domenech P, Reed MB (2009) Rapid and spontaneous loss of phthiocerol dimycocerosate (PDIM) from Mycobacterium tuberculosis grown in vitro: implications for virulence studies. Microbiology 155: 3532–3543. doi:10.1099/mic.0.029199-0.

95. Betts JC, Lukey PT, Robb LC, McAdam RA, Duncan K (2002) Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. Mol Microbiol 43: 717–731. doi:10.1046/j.1365-2958.2002.02779.x.

96. Rodriguez JE, Ramirez AS, Salas LP, Helguera-Repetto C, Gonzalez-y-Merchand J, et al. (2013) Transcription of Genes Involved in Sulfolipid and Polyacyltrehalose Biosynthesis of Mycobacterium tuberculosis in Experimental Latent Tuberculosis Infection. PLoS ONE 8. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3589379/. Accessed 10 August 2014.

97. Schnappinger D, Ehrt S, Voskuil MI, Liu Y, Mangan JA, et al. (2003) Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages Insights into the Phagosomal Environment. J Exp Med 198: 693–704. doi:10.1084/jem.20030846.

98. Talaat AM, Lyons R, Howard ST, Johnston SA (2004) The temporal expression profile of Mycobacterium tuberculosis infection in mice. Proc Natl Acad Sci U A 101: 4602–4607. doi:10.1073/pnas.0306023101.

99. Singh A, Gupta R, Vishwakarma RA, Narayanan PR, Paramasivan CN, et al. (2005) Requirement of the mymA Operon for Appropriate Cell Wall Ultrastructure and Persistence of Mycobacterium

tuberculosis in the Spleens of Guinea Pigs. J Bacteriol 187: 4173–4186. doi:10.1128/JB.187.12.4173-4186.2005.

100. Wimberly H, Shee C, Thornton PC, Sivaramakrishnan P, Rosenberg SM, et al. (2013) R-loops and nicks initiate DNA breakage and genome instability in non-growing Escherichia coli. Nat Commun 4. Available: http://www.nature.com/ncomms/2013/130705/ncomms3115/full/ncomms3115.html. Accessed 31 December 2014.

101. Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol 12: R112. doi:10.1186/gb-2011-12-11-r112.

102. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM (2004) Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proc Natl Acad Sci U S A 101: 4871–4876. doi:10.1073/pnas.0305627101.

103. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, et al. (2013) Mycobacterial Lineages Causing Pulmonary and Extrapulmonary Tuberculosis, Ethiopia. Emerg Infect Dis 19: 460–463. doi:10.3201/eid1903.120256.

104. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. Available: http://arxiv.org/abs/1303.3997. Accessed 23 June 2014.

105. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498. doi:10.1038/ng.806.

106. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. doi:10.1093/bioinformatics/btp352.

107. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: pp. 10–12. doi:10.14806/ej.17.1.200.

108. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760. doi:10.1093/bioinformatics/btp324.

109. Reddy TBK, Riley R, Wymore F, Montgomery P, DeCaprio D, et al. (2009) TB database: an integrated platform for tuberculosis research. Nucleic Acids Res 37: D499–D508. doi:10.1093/nar/gkn652.

110. Kofler R, Pandey RV, Schlotterer C (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27: 3435–3436. doi:10.1093/bioinformatics/btr589.

111. R Development Core Team (n.d.) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available: http: //www.R-project.org/.

112. Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 48: 77–84. doi:10.1046/j.1365-2958.2003.03425.x.

113. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics: btu153. doi:10.1093/bioinformatics/btu153.

114. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res 13: 2178–2189. doi:10.1101/gr.1224503.

115. Katoh K, Standley DM (2014) MAFFT: iterative refinement and additional methods. Methods Mol Biol Clifton NJ 1079: 131–146. doi:10.1007/978-1-62703-646-7_8.

116. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics: btp348. doi:10.1093/bioinformatics/btp348.

117. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313. doi:10.1093/bioinformatics/btu033.

118. Huson DH, Scornavacca C (2012) Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. Syst Biol: sys062. doi:10.1093/sysbio/sys062.

Supplementary Files

File name: S1_Figure.eps

File format: .eps

Title: Patterns of nucleotide diversity (π), Watterson's theta (θ), and Tajima's D across the *Mycobacterium tuberculosis* chromosome for Patient B.

Description: Sliding-window analyses were performed using 100-Kb windows with a step-size of 10-Kb on a uniformly subsampled alignment of 25X for each sample of patient B. Chromosomal coordinates reflect the genomic positions of the reference strain H37Rv, against which pooled-sequence reads were mapped. Blue lines correspond to the first temporal sample, while coral lines correspond to the second temporal sample.

File name: S2_Table.xlsx

File format: .xlsx

Title: *Mycobacterial tuberculosis* strains used for global and regional datasets in this study.

Description: Strains used in this study are a subset of those used in a previous study (Comas *et al*. 2013). Information pertaining to the place of birth of the patient, the place of isolation of the strain, and the phylogeographic area are taken from Comas *et al.* (2013) - Supplementary Table 1 and reported here for the ease of the reader. We performed phylogenetic analysis on the selected strains and confirmed the lineages reported by Comas *et al.* 2013. Accession numbers are listed for each strain.

File name: S3_Table.xlsx

File format: .xlsx

Title: Functional enrichment analysis of genes with extreme values of Tajima's D.

Description: Genes with Tajima's D values in the top 5% (tab A) and bottom 5% (tab B) of the distribution of each sample were tested for enrichment of functional categories (described in *methods*) using a two-sided Fisher's exact test. To account for multiple hypotheses testing, a false discovery rate of 5% was used and the resulting *q*-values are reported. Red font and cell highlighting indicates significance at the 0.05 level. Note that the results presented in tab B are visualized in Fig. 2.

File name: S4_Table.xlsx

File format: .xlsx

Title: Gene-by-gene estimates of population genetic parameters from within- and between-host samples.

Description: For 3,541 genes in the H37Rv genome, population genetic parameters estimated for global and within-host samples are displayed. Genes had to be resolved at ≥ 55% (in at least 75% of strains for the between-host dataset) in order for statistics to be calculated. ABREVIATIONS: COV – fraction of gene resolved at sufficient coverage; □ - Watterson's theta; π – nucleotide diversity; TD – Tajima's D; π_N/π_S – nonsynonymous nucleotide diversity per synonymous nucleotide diversity. See *methods* for details of how each parameter was calculated. For TD and π_N/π_S, the top and bottom 5% of each samples distribution is highlighted in green and red, respectively.

File name: S5_Table.xlsx

File format: .xlsx

Title: SNPs with extreme F_{ST} values in serial samples of within-host *Mycobacterium tuberculosis* populations.

Description: Allele frequency change between longitudinal samples, nucleotide change, and amino acid change are with respect to the minor allele of the first sample time point of each patient (not the H37Rv reference). F_{ST} values reported for patient A are the highest of the three pairwise comparisons (a1-a2, a2-a3, a1-a3).

File name: S5_Table.xlsx

File format: .xlsx

Title: Strain names and accession numbers for genomes used in Figure 7.

Description: Strain names and accession numbers are listed for the genomes used to generate the core genome alignment and maximum likelihood tree.

File name: S7_File.pdf

File format: .pdf

Title: Sensitivity analysis and parameter choice justification for PoPoolation software.

Description: Supplementary information on our sensitivity analysis of the PoPoolation Software. This document includes justification of our parameter choices, as well as four figures with legends in the document.