1 Title
2 Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia
3
4 Author Affiliation
5 Mary B O'Neill[a,b], Andrew Kitchen[c], Alex Zarley[d], William Aylward[e], Vegard Eldholm[f],
6 Caitlin S Pepperell[b,g]
7
8 [a]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA
9 [b]Department of Medical Microbiology and Immunology, University of Wisconsin-Madison, Madison, WI
10 53706, USA
11 [c]Department of Anthropology, University of Iowa, Iowa City, IA 52242, USA
12 [d]Department of Geography, University of Wisconsin-Madison, WI 53706, USA
13 [e]Department of Classical and Ancient Near Eastern Studies, University of Wisconsin-Madison, Madison,
14 WI 53706, USA
15 [f]Infection Control and Environmental Health, Norwegian Institute of Public Health, 0456 Oslo, Norway
16 [g]Department of Medicine, University of Wisconsin-Madison, Madison, WI 53706, USA
17
18 Corresponding Author
19 Caitlin S Pepperell
20 1550 Linden Drive
21 5301 Microbial Sciences Building
22 Madison, WI 53706
23 (608) 262-5983
24 cspepper@medicine.wisc.edu
25
26 Keywords
27 phylogeography, evolution, pathogen, migration, demography
28

Abstract

Mycobacterium tuberculosis (*M.tb*) is a globally distributed, obligate pathogen of humans that can be divided into seven clearly defined lineages. How the ancestral clone of *M.tb* spread and differentiated is important for identifying the ecological drivers of the current pandemic. We reconstructed *M.tb* migration in Africa and Eurasia, and investigated lineage specific patterns of spread. Applying evolutionary rates inferred with ancient *M.tb* genome calibration, we link *M.tb* dispersal to historical phenomena that altered patterns of connectivity throughout Africa and Eurasia: trans-Indian Ocean trade in spices and other goods, the Silk Road and its predecessors, the expansion of the Roman Empire and the European Age of Exploration. We find that Eastern Africa and Southeast Asia have been critical in the dispersal of *M.tb*. Our results reveal complex relationships between spatial dispersal and expansion of *M.tb* populations, and delineate the independent evolutionary trajectories of bacterial sub-populations underlying the current pandemic.

Introduction

The history of tuberculosis (TB) has been rewritten several times as genetic data accumulate from its causative agent, *Mycobacterium tuberculosis* (*M.tb*). In the nascent genomic era, these data refuted the long-held hypothesis that human-adapted *M.tb* emerged from an animal adapted genetic background represented among extant bacteria by *Mycobacterium bovis*, another member of the *Mycobacterium tuberculosis* complex (MTBC) (1). Genetic data from bacteria infecting multiple species of hosts revealed that animal-adapted strains form a nested clade within the diversity of extant *M.tb* (1–3).

M.tb can be classified into seven well-differentiated lineages, which differ in their geographic distribution and association with human sub-populations (4, 5). This observation led to the hypothesis that *M.tb* diversity has been shaped by human migrations out of Africa, and that the most recent common ancestor (MRCA) of extant *M.tb* emerged in Africa approximately 73,000 years ago (6). Human out-of-Africa migrations are a plausible means by which *M.tb* could have spread globally. However, several features of *M.tb* population genetics suggest it has diversified over relatively short time scales: for example, the species is characterized by low genetic diversity (7, 8) and high rates of non-synonymous polymorphism (9).

The observation of limited diversity among extant *M.tb* could be reconciled with the out-of-Africa scenario if *M.tb*'s rate of evolution were orders of magnitude lower than estimates from other bacterial pathogens (6, 8). There are at least nine published estimates of the rate of *M.tb* molecular evolution, which use a variety of calibration methods (10). Rate estimates calibrated with sampling dates, historical events, experimental infection in non-human primates, recent transmission events, and ancient DNA are concordant, and demonstrate expected rate decay as the window of observation shifts from epidemiological to historical time scales. Critically, these *M.tb* rate estimates are similar to those of other bacterial species, and are inconsistent with the out of Africa hypothesis.

When calibrated with ancient DNA, the estimates of the time to most recent common ancestor (TMRCA) for the MTBC are <6,000 years before present (11, 12). This is not necessarily the time period over which TB first emerged, as it is possible – particularly given the apparent absence of recombination among *M.tb* (13) – that the global population has undergone clonal replacement events that displaced ancient diversity from the species.

M.tb is an obligate pathogen of humans with a global geographic range. The finding of a recent origin for the extant *M.tb* population raises the question of how the organism could have spread within this timeframe to occupy its current distribution. *M.tb* populations in the Americas show the impacts of European colonial movements as well as recent immigration [e.g. (14)]; the role of other historical phenomena in driving TB dispersal is not well understood. Here we sought to reconstruct the migratory history of *M.tb* populations in Africa and Eurasia within the newly established framework of a recent origin and evolutionary rates derived from ancient DNA data (11, 12). We discovered lineage-specific patterns of migration and a complex relationship between *M.tb* population growth and migration. Our results connect *M.tb* migration to major historical events in human history that altered patterns of connectivity in Africa and Eurasia. These findings provide context for a recent evolutionary origin of the MRCA of *M.tb* (11–13), which represents yet another paradigm shift in our understanding of the history and origin of this successful pathogen.

Results

A maximum likelihood phylogeny inferred from whole genome sequence (WGS) data from 552 globally extant isolates is shown in Fig. S1. This shows the well described *M.tb* lineage structure, and some associations are evident between lineages and geographic regions (defined here by the United Nations geoscheme). Geographic structure is most evident for lineages 5 and 6 (L5 and L6), limited to West Africa, and lineage 7 (L7), which is restricted to Ethiopia. The phylogeny has an unbalanced shape, with long internal branches that define the lineages and feathery tips, consistent with recent population expansion across lineages.

Heatmaps of lineage frequencies, based on spoligotyping data from 42,358 isolates, are shown in Fig. 1. Geographic patterns in prevalence vary between lineages. Lineage 1 (L1) is prevalent in regions bordering the Indian Ocean, extending from Eastern Africa to Melanesia. Lineage 2 (L2) is broadly distributed, with a predominance in Eastern Eurasia and South East Asia. Lineage 3 (L3) is similar to L1 in that its distribution rings the Indian Ocean, but it does not extend into Southeastern Asia, it has a stronger presence in Northern Africa, and a broader distribution across Southern Asia. Lineage 4 (L4) is strikingly well dispersed, with a predominance throughout Africa and Europe and the entire region bordering the Mediterranean. L5 and L6 are found at low frequencies in Western and Northern Africa. L7, as previously described, is limited to Ethiopia.

Summary statistics comparing the complete sample and individual lineages are shown in Table 1. Genetic diversity, as measured by the numbers of segregating sites and

pairwise differences (Watterson's Θ and π), varied among lineages. L1 and L4 group together and have the highest diversity; L2, L3, L5, and L6 have similar levels of diversity and form the middle grouping; L7 has the lowest diversity. We used the methods implemented in ∂a∂i to reconstruct the demographic histories of *M.tb* lineages from their synonymous site frequency spectra (SFS). Expansion offered an improved fit to the data in comparison with the constant population size model for each dataset (Table 1, Fig. S2).

Individual phylogenetic structures vary among lineages (Fig. 2), likely reflecting their distinct demographic histories. Branch lengths are relatively even across the phylogenies of L1 and L4, whereas L2 and L3 have a less balanced phylogenetic structure. The long, sparse internal branches and radiating tips of L2 and L3 phylogenies are consistent with an early history during which the effective population size remained small (and diversity was lost to drift), followed by more recent population expansion. L5 has a star-like structure, consistent with rapid population expansion.

We used an analysis of molecular variance (AMOVA) to delineate the effects of population sub-division on *M.tb* diversity (Table 1). The global population was highly structured among UN subregions (21% of variation attributable to between-region comparisons), whereas this structure was less apparent when regions were defined by botanical continents (14%). This is consistent with *M.tb*'s niche as an obligate human pathogen, with bacterial population structure directly shaped by that of its host population rather than climatic and other environmental features. We obtained similar results when the lineages were considered separately, except for L4, which had little evidence of population structure (4% variation among UN subregions, 2% among botanical continents).

There was evidence of isolation by distance in the global *M.tb* population, as assessed with a Mantel test of correlations between genetic and geographic distances. We defined geographic distances using three schemes: great circle distances, great circle distances through waypoints of human migration (as described in (15)), and distances along historical trade routes. To allow comparisons between the schemes, values were centered and standardized (*see Methods*). Values of the Mantel test statistic were similar for great circle distances (r = 0.16) and trade network distances (r = 0.16), with distances through waypoints reflective of human migration patterns having a lower value (r = 0.14, $p = 0.0001$ for all three analyses). In analyses of human genetic data, adjustment of great circle distances with waypoints results in a higher correlation between genetic and geographic distances (15). Our Mantel test results therefore do not support a pattern of isolation by distance as expected if the original out-of-Africa human migrations were the primary influence on global diversity of extant *M.tb* (6). To further investigate a potential influence of ancient human migration on *M.tb* evolution, we calculated the correlation between *M.tb* genetic diversity (π) within subregions and their average distances from Addis Ababa (great circle distances through waypoints). Contrary to what is observed for human population diversity (15), we did not observe a significant decline in *M.tb* diversity as a function of distance (adjusted R-squared = -0.1, $p = 0.88$), even when samples from the Americas were included (adjusted R-squared =

8.9e-4, p = 0.34, Fig.S3, Dataset S1). We additionally find no trend in diversity as a function of distance at the lineage specific level (Fig. S3, Dataset S1).

The latter result conflicts with a previously published report (16). We note that samples from the Americas consist solely of L4 isolates in the study by Comas *et al.*, unlike other regions in their analysis that contain isolates from multiple lineages. Ethiopia was over-represented relative to other regions in the previously published analysis, and diversity was calculated for isolates spanning very large geographic regions. The greater sample size, finer geographic resolution, and the use of waypoints of human migration in the present analysis may explain our differing results.

We used the methods implemented in BEAST to reconstruct the migratory history of global *M.tb* (Fig. 3, Fig. S4). Using an evolutionary rate calibrated with 18th century *M.tb* DNA (11), which is similar to the rate inferred with data from 1,000 year old specimens (12), our estimate of the time to most recent common ancestor for extant *M.tb* is between 4032 BCE and 2172 BCE (Table 1; date ranges are based on the upper and lower limits of the 95% highest posterior density (HPD) for the rate reported in Kay *et al* which is more conservative than the 95% HPD of the model). We infer an African origin for the MRCA (Eastern or Western subregion, Table 1). Shortly after emergence of the common ancestor, we infer a migration of the L1-L2-L3-L4-L7 ancestor lineage from Western to Eastern Africa (prior to 2683 BCE), with subsequent migrations occurring out of East Africa.

Emergence of L1 follows migration from East Africa to South Asia at some time between the 3rd millennium and 4th century BCE. L1 has an 'out of India' phylogeographic pattern (Fig. S5), with diverse Indian lineages interspersed throughout the phylogeny. This suggests that the current distribution of L1 around the Indian Ocean arose from migrations out of India, from a pool of bacterial lineages that diversified following migration from East Africa.

The phylogeographic reconstruction indicates that following the divergence of L1, *M.tb* continued to diversify in East Africa, with emergence of L7 there, followed by L4 (Table 1, Fig. 3, Fig. S4). The contemporary distribution of L4 is extremely broad (Fig. 1) and in this analysis of the full alignment we infer an East African location for the internal branches of L4. Notably, in the lineage-specific analyses, we infer a European location for these branches (Fig. S6). The difference is likely due to the fact that inference is informed by deeper as well as descendant nodes in the full alignment. Together, these results imply close ties between Europe and Africa during the early history of this lineage with further migration following emergence of the MRCA in the 1st century CE.

After the emergence of L1 and L7 from East Africa, a migration occurring between 697 BCE and 520 CE established L3 in Southern Asia, with subsequent dispersal out of Southern Asia into its present distribution, which includes East Africa (i.e., a back migration of L3 to Africa). We estimate that L2 diversified in Southeast Asia following migration from East Africa at some point between 697 BCE and 20 BCE. A previously

published analysis of L2 phylogeography also inferred a Southeast Asian origin for the lineage (17).

Bayesian skyline plot (BSP) reconstructions of historical population size were consistent with analyses using ∂a∂i and indicated that lineages 1-6 have undergone expansion (Fig. 4, Fig. S7, Table 1). L2 and L3 underwent abrupt expansion at approximately the same time, whereas expansions of L1 and L4 appeared relatively smooth.

Temporal trends in migration also varied among lineages (Fig. 4). L1 was characterized by high levels of migration until approximately the 7th century CE, when the rate of migration decreased abruptly and remained stable thereafter. L3, by contrast, exhibited consistently low rates of migration. L2 and L4 had more variable trends in migration, as each underwent punctuated increases in migration rate. We estimate this took place in the 9th century for L4, whereas the shift occurred later (13th century) in L2. Temporal trends in growth and migration are congruent for L2 and L4, with increases in migration rate preceding population expansions. Taken together, these results suggest that L1 and L3 populations (as well as L5, L6, and L7) grew *in situ*, whereas range expansion may have contributed to the growth of L2 and L4.

A map showing patterns of connectivity among UN subregions and relative rates of *M.tb* migration with strong posterior support is shown in Fig. 5. Southeast Asia was the most connected region, with significant rates of migration connecting it to eight other regions. Eastern Africa, Eastern Europe, and Southern Asia were also highly connected, with significant rates with six, six, and five other regions, respectively. Western Africa, Eastern Asia, and Western Asia were the least connected regions, with just one significant connection each (to Eastern Africa, Southeast Asia, and Eastern Europe, respectively). Our sample from Western Asia is, however, limited (Dataset S2), and migration from this region may have consequently been underestimated. The highest rates of migration were seen between Eastern Asia and Southeastern Asia, and between Eastern Africa and Southern Asia.

Lineage specific analyses showed that migration between Southern Asia, Eastern Africa, and Southeast Asia has been important for the dispersal of L1, whereas Southeast Asia and Eastern Europe have been important for L2 (Fig. S8). L3 is similar to L1 in that there is evidence of relatively high rates of migration between Southern Asia and Eastern Africa. There is also evidence of migration within Africa between the eastern and southern subregions. In the analyses of migration for L4, Eastern Africa appeared highly connected with other regions.

Discussion

Our reconstructions of *M.tb* dispersal throughout the Old World delineate a complex migratory history that varies substantially between bacterial lineages. Patterns of diversity among extant *M.tb* suggest that historical pathogen populations were capable of moving fluidly over vast distances. Using evolutionary rate estimates from ancient

DNA calibration, we time the dispersal of *M.tb* to a historical period of exploration, trade, and increased connectivity among regions of the Old World.

Consistent with prior reports (6), we infer an origin of *M.tb* on the African continent. There is a modest preference for Western Africa over Eastern Africa, likely due to the early branching West African lineages (i.e. *Mycobacterium africanum*, L5 and L6). Larger samples may allow more precise localization of the *M.tb* MRCA, and Northern Africa in particular is under-studied.

We find that the first lineage to have emerged out of Africa was L1, which is currently concentrated in regions bordering the Indian Ocean from East Africa to Melanesia. The genesis of this lineage traces to migration from East Africa to Southern Asia between the 3rd millennium and 4th century BCE, with subsequent dispersal occurring out of the Indian subcontinent. Our phylogeographic analysis suggests that the early history of L1 was characterized by high levels of migration, particularly between Southern Asia and East Africa, and between Southern Asia and Southeast Asia (Fig. 4, Fig. S8). The geographic distribution of L1, the timing of its emergence and spread, as well as patterns of connectivity underlying its dispersal, are all consistent with migration via established trans-Indian Ocean trade routes linking East Africa to Southern and Southeast Asia (Fig. 6). The initial migration overlaps with the so-called Middle Asian Interaction sphere in The Age of Integration (2600-1900 BCE), which is marked by increased cultural exchange and bulk trade in obsidian, metals, shell, precious stones, spices, and aromatics, among other commodities, along a chain of regional networks between civilizations of Egypt, Mesopotamia, the Arabian peninsula, and the Indus Valley (18–22). East-West contact and trade across the Indian Ocean intensified in the first millennium BCE, when primary sources from Egyptian, Greek, and Roman civilization attest to the extension of these maritime networks to include the eastern Mediterranean, the Red Sea, and the Black Sea (23–26). Historical data from the Roman era indicate that crews on trading ships crossing the Indian Ocean comprised fluid assemblages of individuals from diverse regions, brought together under conditions favorable for the transmission of TB (27–30). These ships would have been an efficient means of spreading *M.tb* among the distant regions involved in trade.

L2 may similarly have an origin in East-West maritime trade across the Indian Ocean, as we infer it arose from a migration event from East Africa to Southeast Asia during the 1st millennium BCE. Trade in spices and aromatics, as well as metals and precious stones, continued to proliferate in these regions at this time, and increased sophistication in ship technology allowed for longer voyages (21, 22, 24, 28, 31, 32). L2 appears to have spread out of Southeast Asia, a highly connected region in our analyses of *M.tb* migration, and is currently found across Eastern Eurasia and throughout Southeast Asia (Fig. 1, Fig. 3, Fig. S8). Interestingly, although L2 is dominant in East Asia, the region did not appear to have played a prominent role in dispersal of this lineage, except in its exchanges with Southeast Asia.

In contrast to L1 and L2, L3 appears to have had relatively low rates of migration throughout its history (Fig. 4). The contemporary geographic range of L3 is also

narrower, extending east from Northern Africa, through Western Asia to the Indian subcontinent (Fig. 1). A study of lineage prevalence in Ethiopia showed that L3 is currently concentrated in the north of the country (16), consistent with our observed north to south gradient in its distribution on the African continent. This is in opposition to L1, which has a southern predominance in Ethiopia and across Eastern Africa (Fig. 1). We estimate L3 emerged in South Asia ca. 520 CE (177-739 CE). Pakistan harbors diverse strains belonging to L3 (Fig. S9), and the Southern Asia region was highly connected with Eastern Africa in our analyses (Fig. S8). Trade along the Silk Road connecting Europe and Asia was very active at the time L3 emerged (33, 34) and its distribution suggests it spread primarily along trading routes connecting Northeast Africa, Western Asia and South Asia (30, 33–35) (Fig. 6). We speculate that this occurred *via* overland routes, which may have limited L3's migration relative to maritime dispersal of the other lineages.

The geographic distribution of L4 is strikingly broad (Fig. 1) and it exhibits minimal population structure (Table 1). This suggests L4 dispersed efficiently and continued to mix fluidly among regions, a pattern we would expect if it was carried by an exceptionally mobile population of hosts. L4 is currently concentrated in regions bordering the Mediterranean, and elsewhere throughout Africa and Europe. We estimate the MRCA of L4 emerged in the 1st century CE (range 368 BCE-362 CE), during the peak of Roman Imperial power across the entire Mediterranean world and expansionist Roman policies into Africa, Europe, and Mesopotamia (36, 37). Rome's population eclipsed one million at this time, a first for any city on Earth. This period also corresponds with the zenith of the Roman army, which numbered over 250,000 soldiers annually for more than three centuries (ca. 27 BCE - 305 CE), and peaked at about 450,000 soldiers in the early third century CE. Swift movement over long distances were hallmarks of the army, as were stations near the frontiers of the vast empire, where soldiers intermarried with locals and veterans settled (35). The empire reached its greatest territorial extent in the early second century CE, when all of North Africa, from the Atlantic Ocean to the Red Sea, was under a single power, with trade on land and sea facilitated by networks of stone-paved roads and protected maritime routes (33, 36, 39). Primary sources from Roman civilization attest to trade with China, purposeful expeditions for exploration, cartography and trade in the Red Sea and Indian Ocean (25, 29, 40–42).

We hypothesize that the broad distribution of L4 reflects rapid diffusion from the Mediterranean region along trade routes extending throughout Africa, the Middle East, and on to India, China, and Southeast Asia. High rates of migration appear to have been maintained for this lineage over much of its evolutionary history (Fig. 4); patterns of connectivity implicate Europe and Africa in its dispersal (Fig. S8). The association of L4 with European migrants is well described, particularly migrants to the Americas (5, 14). We also note the origin and concentration of this lineage on the African continent. Our sample of L4 isolates includes several deeply rooting African isolates, and African isolates are interspersed throughout the phylogeny (Fig. S6). Our analyses showed an increase in migration ca. 9th century: given the distribution of L4, the Arab slave trade is a plausible contributing factor explaining the timing of this increase (28, 43, 44).

The migratory histories of L5, L6, and L7 are less complicated than those of lineages 1-4. Specifically, L5 and L6 are restricted to Western Africa and L7 is found only in Ethiopia (Fig. 3). The reasons for the restricted distributions of these lineages are not immediately obvious: there is evidence in our analyses that other lineages migrated in and out of Western Africa, and Eastern Africa emerged as highly connected and central to the dispersal of *M.tb*. A potential explanation is restriction of the pathogen population to human sub-populations with distinct patterns of mobility and connectivity that did not facilitate dispersal. This is likely the case for L7, which was discovered only recently (45), and is currently largely restricted to the highlands of northern Ethiopia (16). In the case of L6 (also known as *Mycobacterium africanum*), there is evidence suggesting infection is less likely to progress to active disease than for *M. tuberculosis sensu stricto* (46), which could have played a role in limiting its dispersal.

Our reconstructions of *M.tb*'s migratory history suggest that patterns of migration were highly dynamic: the pathogen appears to have dispersed efficiently, in complex patterns that nonetheless preserved the distinct structure of each lineage. Some findings, notably population expansion, were consistent across lineages. Though growth of the global *M.tb* population has been described previously (6, 13), we found here that the pace and magnitude of expansion, and its apparent relationship to trends in migration, varied among lineages (Fig. 4, Fig. S7). Interestingly, L2 and L3 share a similar phylogenetic structure (Fig. 2) and we estimate they have both undergone large-scale, similarly timed (ca. 13th century) expansions (Fig. 4).

The expansion of L2 was preceded by an impressive increase in its rate of migration, suggesting that growth of the pathogen population was facilitated by expansion into new niches. Our phylogeographic reconstructions implicate Russia, Central and Western Asia in L2's recent migratory history (Fig. S8, Fig. S10), which is consistent with a recently reported phylogeographic analysis of L2 (17). The inferred timing of L2's growth and increased migration is close to the well documented incursion of *Yersinia pestis* from Central Asia into Europe that resulted in explosive plague epidemics (47). The experience with plague suggests that patterns of connectivity among humans and other disease vectors were shifting at this place and time, which would potentially open new niches for pathogens including *M.tb*. A similar pattern of growth preceded by geographic range expansion is seen in L4, in the 9th and 15th centuries. The latter expansion coincides with the onset of the 'age of exploration' (48), which would have provided numerous opportunities for spread of this lineage from Europeans to other populations. L1 underwent a similarly timed expansion (Fig. 4) but in this case it appears to have grown *in situ*, e.g. due to changing ecological conditions and/or growth of local human populations. A study of the molecular epidemiology of TB in Vietnam identified numerous recent migrations of L2 and L4 into the region, versus a stable presence of L1 (49); this is consistent with our finding of higher recent rates of migration for L2 and L4 versus L1. A pattern similar to L1 has been identified previously, in the delay between dispersal of *M.tb* from European migrants to Canadian First Nations and later epidemics of TB driven by shifting disease ecology (14). L1 and L4 are similar in that these lineages with a long history of high migration rates underwent smooth

395 expansions, versus the abrupt expansions observed in the setting of a consistently low
396 migration rate (L3), or recent, abrupt increase in migration (L2). These results
397 demonstrate the complex relationship between *M.tb* population growth and migration,
398 and show that under favorable conditions the pathogen can expand into novel niches or
399 accommodate growth in an existing niche.
400
401 While this study comprises the largest phylogeographic analysis on *M.tb* done to date,
402 with 552 isolates collected from 51 countries and all seven described lineages
403 represented, it has some important limitations. We did not attempt to estimate the rate
404 or timescale of *M.tb* evolution, instead relying on published rates that were calibrated
405 with ancient DNA. This is an active area of research, and newly discovered ancient *M.tb*
406 DNA samples will likely refine inference of both the timing and locations of historical
407 migration events, though it is critical to note that recent substitution rate estimates of
408 *M.tb* have converged on rates around $5x10^{-8}$ substitutions per site per year (10). The
409 addition of contemporary isolates of *M.tb* from under-represented regions of the globe
410 would additionally strengthen the sample. While thousands of *M.tb* genomes have been
411 sequenced to date, certain geographic locations are under-studied. For example, we
412 were unable to obtain WGS data for isolates from Northern and Middle Africa. Samples
413 from under-studied regions may allow more precise localization of the MRCA of *M.tb*
414 and uncover new connections that were important in its dispersal throughout the globe.
415
416 Methods
417 **Sample Description.** We assembled/aligned publicly available whole genome
418 sequences (WGS) of *M.tb* isolates for which country of origin information were known
419 and corresponded to traditional definitions of the Old World. We excluded countries
420 where the majority of TB cases are identified in immigrants (50–55). Numbers of
421 isolates per country were selected based on the availability of appropriate genome
422 sequence data as well as relative TB prevalence (56). Our sample necessarily contains
423 a large number of drug-resistant isolates as these are more commonly sequenced.
424 Countries with large numbers of available genomes were sub-sampled such that
425 phylogenetic lineage diversity was captured: phylogenetic inference on all isolates
426 available from a country was performed with Fasttree (57) and a random isolate was
427 selected from each clade extending from *n* branches, where *n* was the desired number
428 of isolates from the country. We also included all isolates belonging to lineages 5-7. Our
429 final dataset consisted of the WGS of 552 previously published *M.tb* isolates collected
430 from 51 countries spanning 13 UN geoscheme subregions. Accession numbers and
431 pertinent information about each sample can be found in Dataset S2. Isolates were
432 assembled via reference guided assembly (RGA) when FASTQ data were available and
433 by multiple genome alignment (MGA) when only draft genome assemblies were
434 accessible.
435
436 **Reference Guided Assembly.** Previously published FASTQ data were retrieved from
437 the National Center for Biotechnology Information (NCBI) sequence read archive (SRA)
438 (58). Low-quality bases were trimmed using a threshold quality of 15, and reads
439 resulting in less than 20bp length were discarded using Trim Galore!
440 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which is a wrapper

tool around Cutadapt (59) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped to H37Rv (NC_000962.3) (60) with the MEM algorithm (61). Duplicates were removed using Picard Tools (http://picard.sourceforge.net), and local realignment was performed with GATK (62). To ensure only high quality sequencing data were included, individual sequencing runs for which <80% of the H37Rv genome was covered by at least 20X coverage were discarded, as were runs for which <70% of the reads mapped as determined by Qualimap (63). Pilon (64) was used to call variants with the following parameters: --variant --mindepth 10 --minmq 40 --minqual 20.

Multiple Genome Alignment. Draft genome assemblies were aligned to H37Rv (NC_000962.3) (60) with Mugsy v1.2.3 (65). Regions not present in H37Rv were removed and merged with the reference-guided assembly.

SNP alignment. VCFs were converted to FASTAs with in-house scripts that treat ambiguous calls and deletions as missing data (available at https://github.com/pepperell-lab/RGAPepPipe). Transposable elements, phage elements, and repetitive families of genes (PE, PPE, and PE-PGRS gene families) that are poorly resolved with short read sequencing were masked to missing data. Isolates with >20% missing sites were excluded from the final Old World collection (Dataset S1). Variant positions with respect to H37Rv were extracted with SNP-sites (66) resulting in 60,818 variant sites. Only sites were at least half of the isolates had confident data (i.e., non-missing) were included in the phylogeographic models and population genetic analyses (60,787 variant sites; 3,838,249 bp). Only 1.7% of variant sites landed in loci associated with drug resistance (Table S1).

Geographic Information. Geographic locations for each of the 552 samples in the Old World collection were obtained from NCBI and/or the respective publications from which the isolates were first described. When precise geographic information was available (e.g., city, province, etc.), coordinates were obtained from www.mapcoordinates.net. When only country level geographic information was found, the 'Create Random Point' tool in ArcGIS 10.3 was used to randomly place each isolate without specific latitude and longitude inside its respective country; inhospitable areas (e.g., deserts and high mountains) and unpopulated areas from each country using 50m data from Natural Earth (http://www.naturalearthdata.com/downloads, accessed February 17, 2016) were excluded as possible coordinates. The 'precision' column of Dataset S1 reflects which method was used.

Trade Route Information. Data for all trade routes active throughout Europe, Africa, and Asia by 1400 CE were compiled from the Old World Trade Routes (OWTRAD) Project (www.ciolek.com/owtrad.html, accessed February 17, 2016). For each route, both node information (trade cities, oases, and caravanserai) and arc information (the routes between nodes) were imported into ArcGIS (Fig. 6). *M.tb* isolate locations were also imported as points and the 'Generate Near Table' tool was used to assign each isolate to its nearest node in the trade network and is listed in the 'NearPost' column of Dataset S2.

Maximum Likelihood Inference. We used RAxML v8.2.3 (67) for maximum likelihood phylogenetic analysis of the Old World *M.tb* alignment (all sites where at least half of isolates had non-missing data) under the general time reversible model of nucleotide substitution with a gamma distribution to account for site-specific rate heterogeneity. Rapid bootstrapping of the corresponding SNP alignment was performed with the -autoMR flag, converging after 50 replicates. Tree visualization was created with the ggtree package in R (68).

Lineage Frequencies. The SITVIT WEB database (69), which is an open access *M.tb* molecular markers database, was accessed on September 5, 2016. Spoligotypes were translated to lineages based on the following study (70). The following conversions were also included: EAI7-BGD2 for L1, CAS for L3, and LAM7-TUR, LAM12-Madrid1, T5, T3-OSA, and H4 for L4. Isolates containing ambiguous spoligotypes (denoted with >1 spoligotype) were inspected manually and assigned to appropriate lineages. Relative lineage frequencies of lineages 1-6 for each country containing data for >10 isolates were calculated and plotted with the rworldmap package in R (71).

Population genetic statistics. Nucleotide diversity (π) and Watterson's theta (Θ) for various population assignments (e.g., lineage, UN subregion) were calculated with EggLib v2.1.10 (72).

Demographic inference from the observed site frequency spectrum (SFS). SNP-sites (66) was used to convert the Old World *M.tb* alignment to a multi-sample VCF and SnpEff (73) was used to annotate variants with respect to H37Rv (NC_000962.3) (60) as synonymous, non-synonymous, or intergenic. Loci at which any sequence in the population had a gap or unknown character where removed from the data set. Demographic inference with the synonymous SFS for each of the seven lineages and the entire collection was performed using ∂a∂i (74). We modeled constant population size (standard neutral model) and an instantaneous expansion model, and identified the best-fit model and maximal likelihood parameters of the demographic model given our observed data. Our parameter estimates, ν and τ, were optimized for the instantaneous expansion and exponential growth models. Uncertainty analysis of these parameters were analyzed using the Godambe Information Matrix (75) on 100 samplings of the observed synonymous SFS with replacement and subsequent model inference.

Phylogeographic & Demographic Inference. The Old World *M.tb* SNP alignment and individual lineage SNP alignments were analyzed using the Bayesian Markov Chain Monte Carlo coalescent method implemented in BEAST v1.8 (76) with the BEAGLE library (77) to facilitate rapid likelihood calculations. Analyses were performed using the general time reversible model of nucleotide substitution with a gamma distribution to account for rate heterogeneity between sites, a strict molecular clock, and both constant and Bayesian skyline plot (BSP) demographic models. Country of origin or the UN subregion for each isolate was modeled as a discrete phylogenetic trait (78). All Markov chains were run for at least 100 million generations, sampled every 10,000 generations, and with the first 10,000,000 states discarded as burn-in; replicate runs were performed

for analyses and combined to assess convergence. Estimated sample size (ESS) values of non-nuisance parameters were >200 for all analyses. Site and substitution model choice were based on previous analyses of *M.tb* global alignments as opposed to an exhaustive comparison of models which would require unreasonable computational resources. Strict vs relaxed molecular clocks did not result in altered trends of migration at the lineage level, and comparisons between analyses using strict and relaxed clocks show strong correlation between the estimated height of nodes (e.g., $R^2 > 0.97$; Fig. S11, Fig. S12). Table S2 provides a summary of BEAST analyses presented and the results derived from them. Tree visualizations were created with FigTree (http://tree.bio.edu.ac.uk/software/figtree/) and the ggtree package in R (68).

We note that phylogeographic inference methods are an active area of research and increasingly sophisticated models are continuously being developed [e.g. (79, 80)]. We found alternative methods unsuitable and/or intractable for our large dataset. As methods improve, comparison of the results inferred herein to other phylogeographic models will be important to investigate the sensitivity of our results to the method of phylogeographic inference.

Analysis of Molecular Variance (AMOVA). AMOVAs were performed using the 'poppr.amova' function (a wrapper for the ade4 package (81) implementation) in the poppr package in R (82). Bins were assigned via the following classification systems: UN geoscheme subregions and Level 1 ('botanical continents') of the World Geographical Scheme for Recording Plant Distributions (WGSRPD). Isolate assignation can be found in Dataset S2. Genetic distances between isolates were calculated with the 'dist.dna' function of the ape v4.0 package in R (83) from the SNP alignment.

Mantel tests. Great circle distances between *M.tb* isolate locations were calculated with the 'distVincentyEllipsoid' function in the geosphere R package (84). Geographic distances between isolate locations along the trade network were calculated by adding the great circle distances from the isolates to the nearest trade hubs and the shortest distance between trade hubs along the trade network; the latter was determined using an Origin-Destination Cost Matrix and the 'Solve' tool in the Network Analyst Toolbox of ArcGIS which calculates the shortest distance from each origin to every destination along the arcs in the trade network. In the event that two isolates were assigned to the same trade post, the great circle distance between the isolates was used. To calculate the geographic distance between isolates in a manner that reflects human migrations, the great circle distance between isolates and waypoints were summed. These were calculated with a custom R function (available at https://github.com/ONeillMB1/Mtb_Phylogeography_v2) using a series of rules to define whether or not the path between isolates would have gone through a waypoint. For all three distance metrics, values were log transformed and standardized. Genetic distances between isolates were calculated with the 'dist.dna' function in the ape v4.0 package in R (83) from the SNP alignment. The 'mantel' function of the vegan package in R (85) was used to perform a Mantel test between the genetic distance matrix and each of the three geographic matrices for both the Old World *M.tb* collection and each

individual lineage. Four of the 552 isolates were excluded from these analyses as they were from Kiribati and trade networks spanning this region were not compiled.

Migration Rate Inference. Migration rates were inferred from the Bayesian maximum clade credibility trees for the entire collection of *M.tb* isolates (*n* = 552). Individual lineages that contain isolates from multiple UN subregions (i.e., L1: *n* = 89, L2: *n* = 181, L3: *n* = *65*, and L4: *n* = 143) were extracted and plotted separately. Only nodes with posterior probabilities greater than or equal to 80% were considered. A migration event was classified as a change in the most probable reconstructed ancestral geographic region from a parent to child node. Median heights of the parent and child nodes were treated as a range of time that the migration event could have occurred. The rate of migration through time for each lineage or the Old World collection was inferred by summing the number of migration events occurring across every year of the time-scaled phylogeny, divided by the number of possible migration events (i.e., the total number of branches in existence during each year of the time-scaled phylogeny). Code for these analyses is available at https://github.com/ONeillMB1/Mtb_Phylogeography_v2.

Additionally, relative migration rates between UN subregions were derived from the BEAST analyses of phylogeography. The Bayesian stochastic search variable selection method (BSSVS) for identifying the most parsimonious migration matrix implemented in BEAST as part of the discrete phylogeographic migration model (78) allowed us to use Bayes Factors to identify the migration rates with the greatest posterior support and provide posterior estimates for their relative rates. Significant relative rates (Bayes Factor > 5) and connectivity among subregions were visualized with Cytoscape v3.2.0 (86) and superimposed onto a map generated with the 'rworldmap' package in R (71).

Relationship between genetic diversity and geographic distance from Addis Ababa. For this analysis, we added Northern and Central American datasets, assembled in an identical manner to those of the Old World collection and masked at sites were less than half of the Old World collection had confident data (3,838,249 bp). For each UN subregion, the mean latitude and longitude coordinates for all *M.tb* isolates within the region were calculated. The great circle distances from these average estimates for regions to Addis Ababa were then calculated, using waypoints for between-continent distance estimates to make them more reflective of presumed human migration patterns (15). Cairo was used as a waypoint for Eastern Europe, Central Asia, Western Asia, Southern Asia, Eastern Asia, and South-Eastern Asia; Cairo and Istanbul were used as waypoints for Western Europe and Southern Europe; Cairo, Anadyr, and Prince Rupert were used as waypoints for Northern and Central America. The distance between each region and Addis Ababa were the sum of the great circle distances between the two points (the average coordinates for the UN subregion and Addis Ababa) and the waypoint(s) in the path connecting them, plus the great circle distance(s) between waypoints if two were used. Treating each UN subregion as a population, the relationship between genetic diversity (assessed with π) and geographic distance from Addis Ababa were explored with linear regression for both the entire Old World collection and individual lineages in R (87). Code is available at https://github.com/ONeillMB1/Mtb_Phylogeography_v2.

Acknowledgments

We would like to thank past and present members of the Pepperell Lab for helpful feedback on the project, and particularly highlight Trent Prall for his contribution to preliminary research on historic human trade routes. We also wish to thank Lucy Weinert for her advice on Mantel test standardization. MBO is supported by the National Science Foundation Graduate Research Fellowship Program (DGE-1255259). CSP is supported by National Institutes of Health (R01AI113287).

References

1. Brosch R, et al. (2002) A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proc Natl Acad Sci* 99(6):3684–3689.

2. Behr MA, et al. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* 284(5419):1520–1523.

3. Hershberg R, et al. (2008) High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. *PLoS Biol* 6(12):e311.

4. 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 A* 101:4871–6.

5. Gagneux S, et al. (2006) Variable host–pathogen compatibility in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* 103(8):2869–2873.

6. Comas I, et al. (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* 45(10):1176–1182.

7. O'Neill MB, Mortimer TD, Pepperell CS (2015) Diversity of Mycobacterium tuberculosis across evolutionary scales. *bioRxiv*:14217.

8. Eldholm V, Balloux F (2016) Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out. *Trends Microbiol* 24(8):637–648.

9. Rocha EPC, et al. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* 239(2):226–235.

10. Eldholm V, et al. (2016) Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis. *Proc Natl Acad Sci* 113(48):13881–13886.

11. Kay GL, et al. (2015) Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun* 6:6717.

12. Bos KI, et al. (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* 514(7523):494–497.

13. Pepperell CS, et al. (2013) The Role of Selection in Shaping Diversity of Natural M. tuberculosis Populations. *PLoS Pathog* 9(8):e1003543.

14. Pepperell CS, et al. (2011) Dispersal of Mycobacterium tuberculosis via the Canadian fur trade. *Proc Natl Acad Sci* 108(16):6526–6531.

15. Ramachandran S, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* 102(44):15942–15947.

16. Comas I, et al. (2015) Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa. *Curr Biol* 25(24):3260–3266.

17. Luo T, et al. (2015) Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese. *Proc Natl Acad Sci* 112(26):8136–8141.

18. Coningham R, Young R (2015) *The Archaeology of South Asia: From the Indus to Asoka, c.6500 BCE–200 CE* (Cambridge University Press) Available at: https://books.google.com/books?id=yaJrCgAAQBAJ.

19. Vogt B (1996) Bronze Age Maritime Trade in the Indian Ocean: Harappan Traits on the Oman Peninsula. *The Indian Ocean in Antiquity*, ed Reade J, pp 107–132.

20. Zarins J (1996) Obsidian in the Larger Context of Predynastic/Archaic Egyptian Red Sea Trade. *The Indian Ocean in Antiquity*, ed Reade J, pp 107–132.

21. Parkin D, Barnes R eds. (2002) *Ships and the development of maritime technology in the Indian Ocean* (RoutledgeCurzon, London).

22. Ray HP (2003) *The archaeology of seafaring in ancient South Asia* (Cambridge University Press, Cambridge ; New York).

23. Boussac M-F, Salles J-F, France eds. (1995) *Athens, Aden, Arikamedu: essays on the interrelations between India, Arabia, and the eastern Mediterranean* (Manohar : Distributed in South Asia by Foundation Books, New Delhi).

24. Ray HP, et al. (1996) *Tradition and archaeology: early maritime contacts in the Indian Ocean* (Manohar Publishers, New Delhi).

25. Dilke O (1985) *Greek and Roman Maps* (Johns Hopkins University Press, Baltimore).

26. Salles J-F (1996) Achaemenid and Hellenistic Trade in the Indian Ocean. *The Indian Ocean in Antiquity*, pp 251–267.

27. Rauh NK (2003) *Merchants, sailors and pirates in the Roman world* (Tempus, Stroud).

28. Wink A (2002) From the Mediterranean to the Indian Ocean: Medieval History in Geographic Perspective. *Comp Stud Soc Hist* 44(3):416–445.

29. Begley V, De Puma RD (1991) *Rome and India: The Ancient Sea Trade* (University of Wisconsin Press, Madison).

30. André J, Filliozat J (1986) *L'Inde vue de Rome: textes latins de l'antiquité, relatifs à l'Inde* (Belles Lettres, Paris).

31. Kent RK (1979) The Possibilities of Indonesian Colonies in Africa with Reference to Madagascar. *Mouvements de Populations Dans L'Ocean Indie* (H. Champion, Paris), pp 93–105.

32. Blench R (1996) The Ethnographic Evidence for Long-distance Contacts between Oceania and East Africa. *The Indian Ocean in Antiquity*, ed Reade J Available at: http://public.eblib.com/choice/publicfullrecord.aspx?p=1517609 [Accessed March 28, 2017].

33. Ball W (2016) *Rome in the East: The Transformation of an Empire* (Routledge, London & New York). 2nd Ed.

34. Hansen V (2012) *The Silk Road: A New History* (Oxford University Press, Oxford).

35. Sartre M (1991) *L'Orient romain: provinces et sociétés provinciales en Méditerranée orientale d'Auguste aux Sévères (31 avant J.-C-235 après J.-C.)* (Seuil, Paris).

36. Luttwak EN (1976) *The grand strategy of the Roman Empire: from the first century A.D. to the third* (Weidenweld & Nicholson, London).

37. Isaac BH (2004) *The limits of empire: the Roman army in the East* (Clarendon Press, Oxford).

38. Whittaker CR (1997) *Frontiers of the Roman Empire: a social and economic study* (The John Hopkins Univ. Press, Baltimore, Md.). Paperback ed.

39. Millar F (1993) *The Roman Near East, 31 B.C.-A.D. 337* (Harvard University Press, Cambridge, Mass).

40. Butcher K (2003) *Roman Syria and the Near East* (J. Paul Getty Museum : Getty Publications, Los Angeles).

41. Erdkamp P (2002) *The Roman Army and the Economy* (Gieben, Amsterdam).

42. Pfister R, Bellinger L (1945) *The Excavations at Dura-Europos. Final Report IV: The Textiles* (Yale University Press, New Haven).

43. Hopper MS (2015) *Slaves of one master: globalization and slavery in Arabia in the age of empire* (Yale University Press, New Haven) Available at: http://dx.doi.org/10.12987/yale/9780300192018.001.0001 [Accessed March 28, 2017].

44. Hourani GF (1995) *Arab seafaring* (Princeton University Press, Princeton, N.J). Expanded ed.

729 45. Blouin Y, et al. (2012) Significance of the Identification in the Horn of Africa of an
730 Exceptionally Deep Branching Mycobacterium tuberculosis Clade. *PLOS ONE* 7(12):e52841.

731 46. Jong D, et al. (2008) Progression to Active Tuberculosis, but Not Transmission, Varies by
732 Mycobacterium tuberculosis Lineage in The Gambia. *J Infect Dis* 198(7):1037–1043.

733 47. Benedictow OJ (2004) *The Black Death, 1346-1353: the complete history* (Boydell Press,
734 Woodbridge, Suffolk, UK ; Rochester, N.Y., USA).

735 48. ʿĀlam M, Subrahmanyam S (2009) *Indo-Persian travels in the age of discoveries, 1400-1800*
736 (Cambridge University Press, Cambridge). Digit. pr.

737 49. Holt KE, et al. (2016) Genomic analysis of Mycobacterium tuberculosis reveals complex
738 etiology of tuberculosis in Vietnam including frequent introduction and transmission of
739 Beijing lineage and positive selection for EsxW Beijing variant. *bioRxiv*:92189.

740 50. Ageing AGD of H and Tuberculosis notifications in Australia, 2008 and 2009. Available at:
741 http://www.health.gov.au/internet/main/publishing.nsf/Content/cda-cdi3601c.htm#refs
742 [Accessed August 2, 2017].

743 51. Government of Canada PHA of C (2005) TUBERCULOSIS PREVENTION AND CONTROL IN
744 CANADA A FEDERAL FRAMEWORK FOR ACTION. Available at: http://www.phac-
745 aspc.gc.ca/index-eng.php [Accessed August 2, 2017].

746 52. White Z, et al. (2017) Immigrant Arrival and Tuberculosis among Large Immigrant- and
747 Refugee-Receiving Countries, 2005–2009. *Tuberc Res Treat*.
748 doi:10.1155/2017/8567893.

749 53. CDC - Reported Tuberculosis in the United States, 2015 - TB Available at:
750 https://www.cdc.gov/tb/statistics/reports/2015/default.htm [Accessed August 2, 2017].

751 54. Tuberculosis in England: annual report - GOV.UK Available at:
752 https://www.gov.uk/government/publications/tuberculosis-in-england-annual-report
753 [Accessed August 2, 2017].

754 55. Tuberculosis in New Zealand: Annual Report 2014 Available at:
755 https://surv.esr.cri.nz/surveillance/AnnualTBReports.php?we_objectID=4251 [Accessed
756 August 2, 2017].

757 56. WHO | Global tuberculosis report 2013 *WHO*. Available at:
758 http://www.who.int/tb/publications/global_report/en/index.html [Accessed November 5,
759 2013].

760 57. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood
761 Trees for Large Alignments. *PLOS ONE* 5(3):e9490.

762 58. Leinonen R, Sugawara H, Shumway M (2011) The Sequence Read Archive. *Nucleic Acids*
763 *Res* 39(Database issue):D19–D21.

764 59. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing
765 reads. *EMBnet.journal* 17(1):10–12.

766 60. Cole ST, et al. (1998) Erratum: Deciphering the biology of Mycobacterium tuberculosis
767 from the complete genome sequence. *Nat Lond* 396(6707):190.

768 61. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-
769 MEM. *ArXiv13033997 Q-Bio*. Available at: http://arxiv.org/abs/1303.3997 [Accessed
770 March 24, 2017].

771 62. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-
772 generation DNA sequencing data. *Nat Genet* 43(5):491–498.

773 63. García-Alcalde F, et al. (2012) Qualimap: evaluating next-generation sequencing alignment
774 data. *Bioinformatics* 28(20):2678–2679.

775 64. Walker BJ, et al. (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant
776 Detection and Genome Assembly Improvement. *PLOS ONE* 9(11):e112963.

777 65. Angiuoli SV, Salzberg SL (2011) Mugsy: fast multiple alignment of closely related whole
778 genomes. *Bioinformatics* 27(3):334–342.

779 66. Page AJ, et al. (2016) *SNP-sites: rapid efficient extraction of SNPs from multi-FASTA*
780 *alignments* Available at: http://biorxiv.org/lookup/doi/10.1101/038190 [Accessed March
781 4, 2016].

782 67. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of
783 large phylogenies. *Bioinformatics* 30(9):1312–1313.

784 68. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y (2017) ggtree: an r package for visualization and
785 annotation of phylogenetic trees with their covariates and other associated data. *Methods*
786 *Ecol Evol* 8(1):28–36.

787 69. Demay C, et al. (2012) SITVITWEB – A publicly available international multimarker
788 database for studying Mycobacterium tuberculosis genetic diversity and molecular
789 epidemiology. *Infect Genet Evol* 12(4):755–766.

790 70. Shabbeer A, et al. (2012) TB-Lineage: An online tool for classification and analysis of strains
791 of Mycobacterium tuberculosis complex. *Infect Genet Evol* 12(4):789–797.

792 71. South A (2016) *rworldmap: Mapping Global Data* Available at: https://cran.r-
793 project.org/web/packages/rworldmap/index.html [Accessed March 24, 2017].

794 72. De Mita S, Siol M (2012) EggLib: processing, analysis and simulation tools for population
795 genetics and genomics. *BMC Genet* 13:27.

796 73. Cingolani P, et al. (2012) A program for annotating and predicting the effects of single
797 nucleotide polymorphisms, SnpEff. *Fly (Austin)* 6(2):80–92.

798 74. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint
799 Demographic History of Multiple Populations from Multidimensional SNP Frequency Data.
800 *PLoS Genet* 5(10):e1000695.

801 75. Coffman AJ, Hsieh PH, Gravel S, Gutenkunst RN (2016) Computationally Efficient
802 Composite Likelihood Statistics for Demographic Inference. *Mol Biol Evol* 33(2):591–593.

803 76. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees.
804 *BMC Evol Biol* 7(1):214.

805 77. Ayres DL, et al. (2012) BEAGLE: An Application Programming Interface and High-
806 Performance Computing Library for Statistical Phylogenetics. *Syst Biol* 61(1):170–173.

807 78. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian Phylogeography Finds
808 Its Roots. *PLOS Comput Biol* 5(9):e1000520.

809 79. De Maio N, Wu C-H, O'Reilly KM, Wilson D (2015) New Routes to Phylogeography: A
810 Bayesian Structured Coalescent Approximation. *PLoS Genet* 11(8):e1005421.

811 80. Lemey P, Rambaut A, Welch JJ, Suchard MA (2010) Phylogeography Takes a Relaxed
812 Random Walk in Continuous Space and Time. *Mol Biol Evol* 27(8):1877–1885.

813 81. Dray S, Dufour A-B, others (2007) The ade4 package: implementing the duality diagram for
814 ecologists. *J Stat Softw* 22(4):1–20.

815 82. Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: an R package for genetic analysis of
816 populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2:e281.

817 83. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R
818 language. *Bioinformatics* 20(2):289–290.

819 84. Hijmans RJ, Williams E, Vennes C (2016) *geosphere: Spherical Trigonometry* Available at:
820 https://cran.r-project.org/web/packages/geosphere/index.html [Accessed March 24,
821 2017].

822 85. Oksanen J, et al. (2017) *vegan: Community Ecology Package* Available at: https://cran.r-
823 project.org/web/packages/vegan/index.html [Accessed March 24, 2017].

824 86. Shannon P, et al. (2003) Cytoscape: A Software Environment for Integrated Models of
825 Biomolecular Interaction Networks. *Genome Res* 13(11):2498–2504.

826 87. R Development Core Team *R: A Language and Environment for Statistical Computing* (R
827 Foundation for Statistical Computing, Vienna, Austria) Available at: http: //www.R-
828 project.org/.

Figures

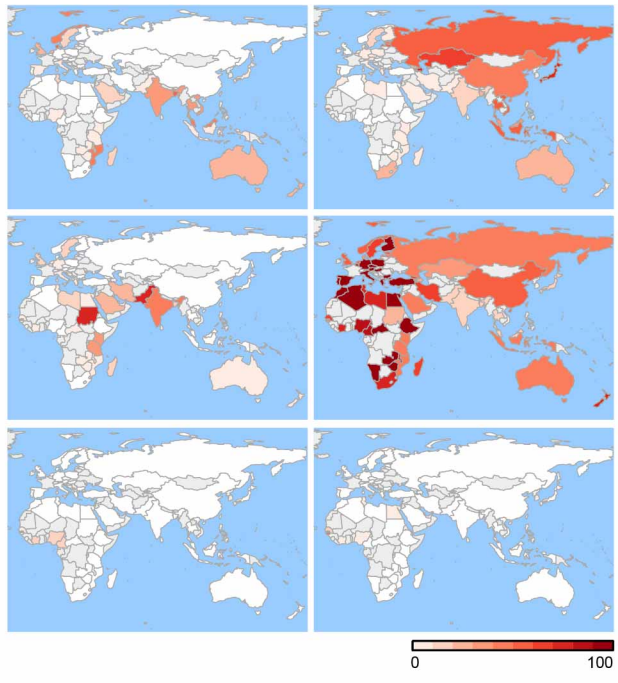


Fig 1. Geographic distributions of *Mycobacterium tuberculosis* lineages 1-6. 42,358 spoligotypes from the SITVIT WEB database were assigned to lineages 1-6 and their relative frequencies calculated in each country with data from > 10 isolates; countries are colored based on the relative proportions of each lineage; unsampled countries are shown in grey.

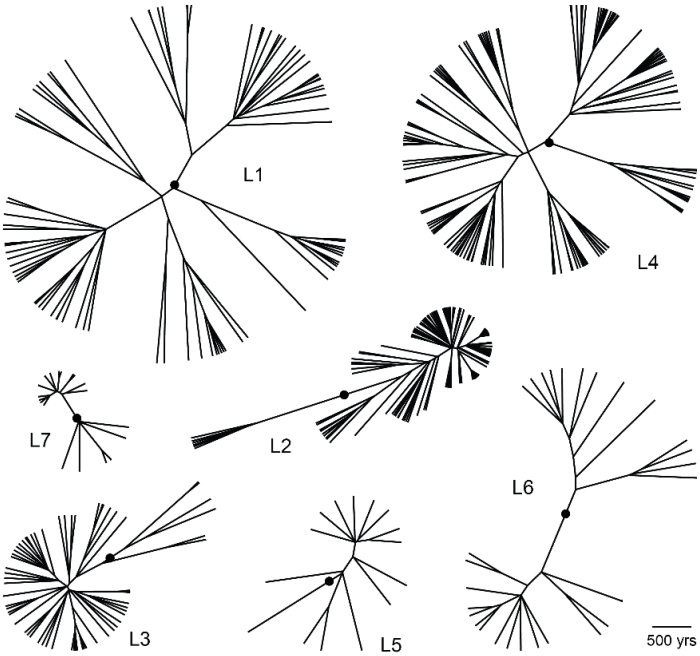


Fig 2. Maximum clade credibility phylogenies of *Mycobacterium tuberculosis* lineages 1-6. Bayesian analyses were performed with the general time reversible model of nucleotide substitution with a gamma distribution to account for rate heterogeneity between sites, a strict molecular clock, and Bayesian skyline plot (BSP) demographic models. The most recent common ancestor (MRCA) of each lineage is indicated with a black circle. Lineage phylogenies were rooted using lineage MRCAs from the phylogeny of the entire Old World collection, which was dated using a substitution rate of 5×10^{-8} substitutions/site/year (11).

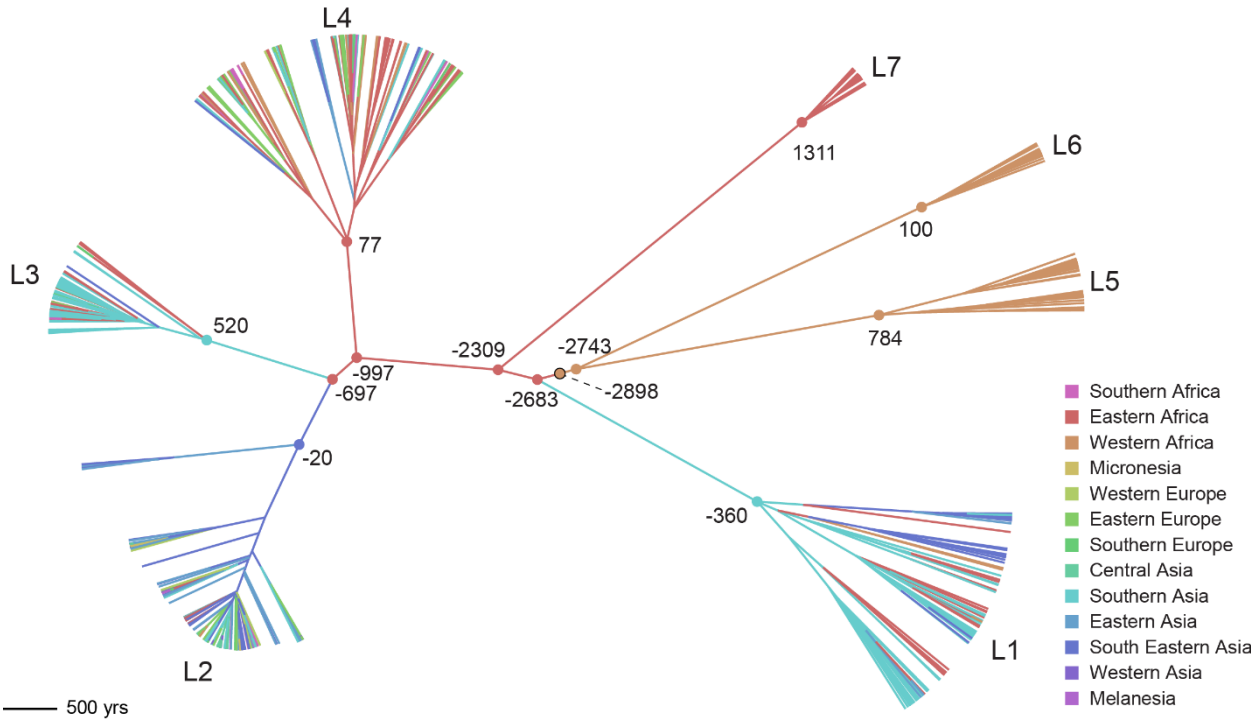


Fig 3. Maximum clade credibility tree of 552 *Mycobacterium tuberculosis* genomes. Estimated divergence dates are shown in calendar years based on median heights and a substitution rate of 5 x 10^{-8} substitutions/site/year (11). Branches are colored according to their most probable subregion in the UN geoscheme. Nodes corresponding to the most recent common ancestors (MRCA) of each lineage, lineage splits, and the MRCA of *M.tb* (outlined black) are marked with circles and colored to reflect their most probable geographic origin.

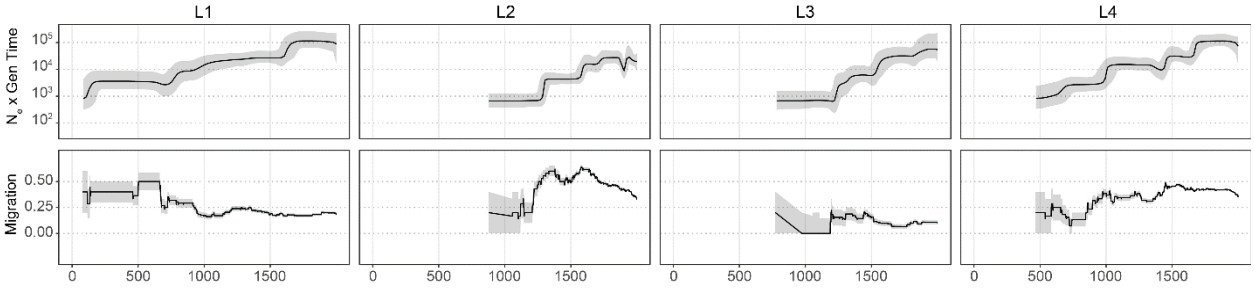


Fig 4. Demographic histories of *Mycobacterium tuberculosis* lineages 1-4. Bayesian skyline plots (BSPs, top panels) show inferred change in effective population time through time from lineage specific analyses. Black lines denote median N_e and gray shading the 95% highest posterior density. Migration events inferred from phylogeographic analysis of the full sample of *M.tb* genomes are shown in the second panels (see *Methods*). Grey shading depicts the rates inferred after the addition or subtraction of a single migration event, and demonstrate the uncertainty of rate estimates, particularly from the early history of each lineage. Dates are shown in calendar years and are based on scaling the phylogeny of the Old World collection with a substitution rate of 5×10^{-8} substitutions/site/year (11).

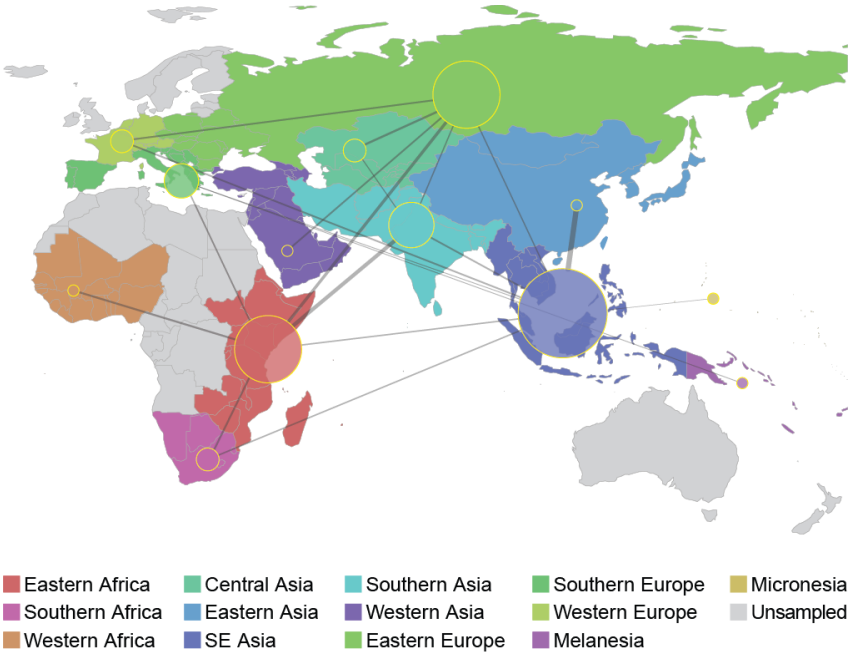


Fig 5. Connectivity of UN regions during dispersal of *M.tb*. The Bayesian stochastic search variable selection method (BSSVS) was used to identify and quantify migrations with strong support in discrete phylogeographic analysis of 552 *M.tb* isolates. Node sizes reflect the number of significant migrations emanating from the region, whereas the thickness of lines connecting regions reflects the relative rate between regions.

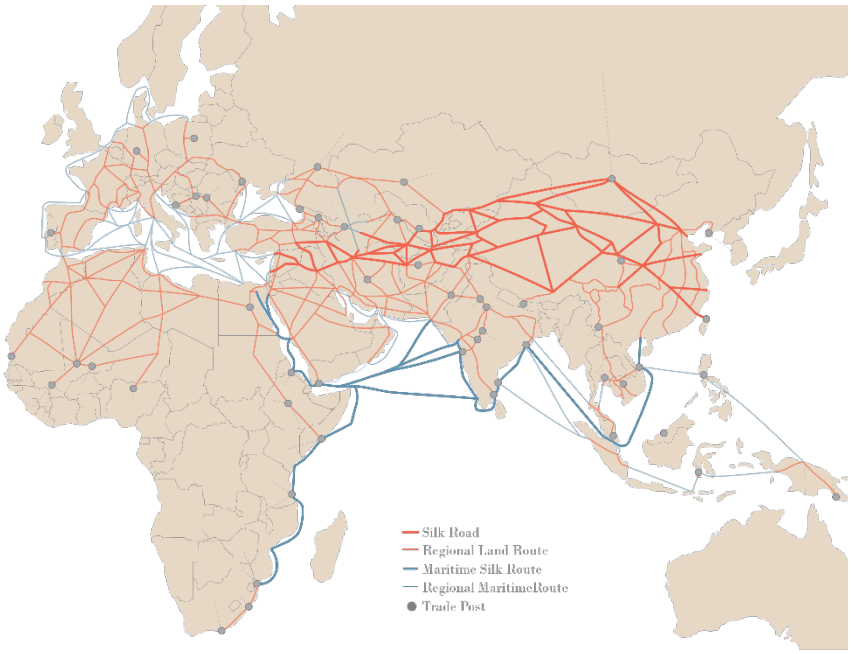
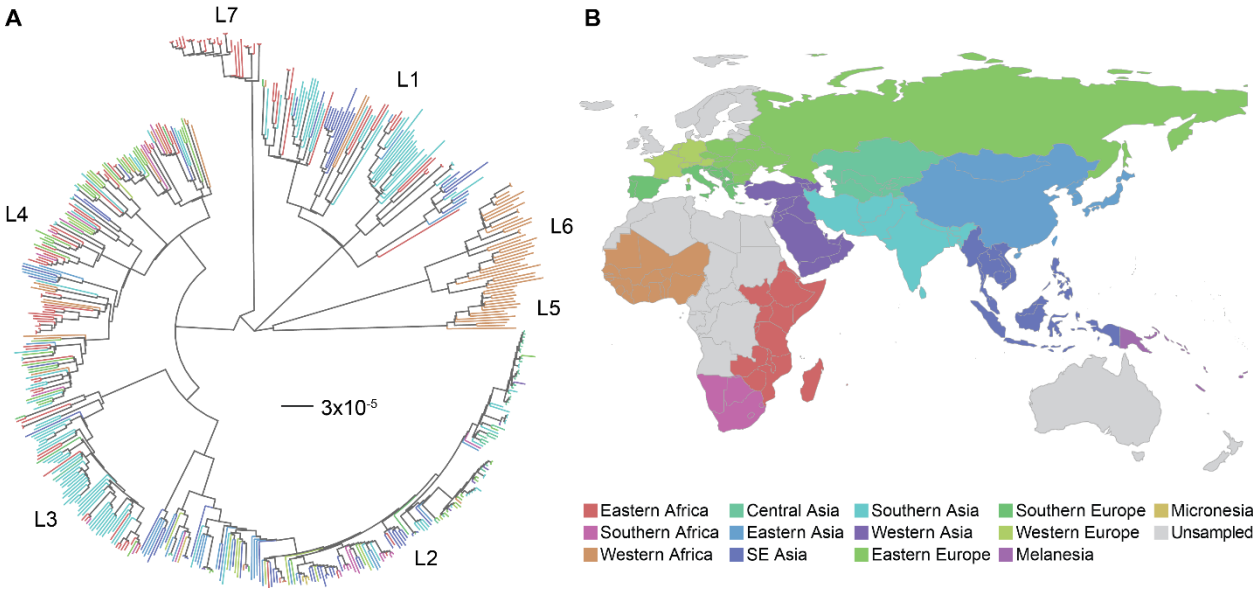


Fig. 6. Trade routes active throughout Europe, Africa and Asia by 1400 CE. Nodes (trade cities, oases, and caravanserai) and arcs (the routes between nodes) are from the Old World Trade Routes (OWTRAD) Project (www.ciolek.com/owtrad.html, accessed February 17, 2016) and are visualized with ArcGIS.

Table 1. Genetic diversity of Old World *M.tb* across lineages 1-7. TMRCA estimates reflect scaling of results to evolutionary rates calibrated from ancient DNA [median 5.00×10^{-8} substitutions/ site/ year (11)]. To account for uncertainty in this rate estimate, our lower and upper TMRCA estimates reflect scaling of our results with the low and high bounds of the 95% highest posterior density estimates of the rate reported from ancient DNA analysis (i.e. 4.06×10^{-8} and 5.87×10^{-8}, respectively).

		MTBC	***L1***	***L4***	***L2***	***L3***	***L5***	***L6***	***L7***
Sample	***n***	552	89	143	181	65	15	31	28
Diversity	**Θ**	2.13E-03	7.56E-04	7.80E-04	4.49E-04	3.88E-04	1.72E-04	3.04E-04	7.99E-05
	π	2.80E-04	1.92E-04	1.54E-04	7.46E-05	9.16E-05	8.77E-05	1.41E-04	4.52E-05
Demographic Inference	***N/Nanc***	91 ± 4	71 ± 5	55 ± 22	112 ± 102	148 ± 2	504 ± 111	50 ± 5	17 ± 4
	Generations (Nanc)	0.16 ± 0.01	0.80 ± 0.06	0.65 ± 0.35	0.41 ± 0.94	3.54 ± 0.04	3.94 ± 0.73	1.10 ± 0.09	2.45 ± 0.89
	LL expansion	-1788.4	-424.2	-492.8	-467.1	-108.2	-42.4	-151.9	-64.5
	LL neutral	-10549.2	-3246.6	-3474.6	-2378.9	-1717.0	-520.7	-912.3	-159.4
	***p*-value**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Structure UN subregions	**Var. Between**	21	19	4	20	16	NA	NA	NA
	Var. Within	79	81	96	80	84	NA	NA	NA
	p-value	<0.001	<0.001	0.001	<0.001	0.004	NA	NA	NA
Structure Botanical Continents	**Var. Between**	14	5	2	9	13	NA	NA	NA
	Var. Within	86	95	98	91	87	NA	NA	NA
	p-value	<0.001	0.02	0.05	<0.001	<0.001	NA	NA	NA
TMRCA	**median**	-2898	-360	77	-20	520	784	100	1311
	lower	-4032	-906	-368	-488	177	502	-339	1152
	upper	-2172	-10	362	279	739	964	382	1413
Geographic origin	**1st region**	W Africa	S Asia	E Africa	SE Asia	S Asia	W Africa	W Africa	E Africa
	probability	54.2%	75.6%	98.9%	81.0%	63.5%	99.9%	99.8%	99.8%
	2nd region	E Africa	E Africa	E Europe	E Asia	E Africa	E Africa	E Africa	S Africa
	probability	37.5%	24.1%	0.7%	9.2%	36.2%	0.1%	0.2%	0.0%

889 Supporting Information
890
891
892



893
894
895 **Fig. S1.** Phylogenetic structure and geographic origin of 552 *Mycobacterium*
896 *tuberculosis* isolates. (*A*) Maximum likelihood phylogeny. Phylogenetic analysis was
897 performed with RAxML using the general time reversible model of nucleotide
898 substitution under the Gamma model of rate heterogeneity. Rapid bootstrapping was
899 performed with the -autoMR flag, converging after 50 replicates. Tip labels are colored
900 with respect to their geographic origin according to the UN geoscheme. (*B*) Map of the
901 Old World colored by UN subregion.

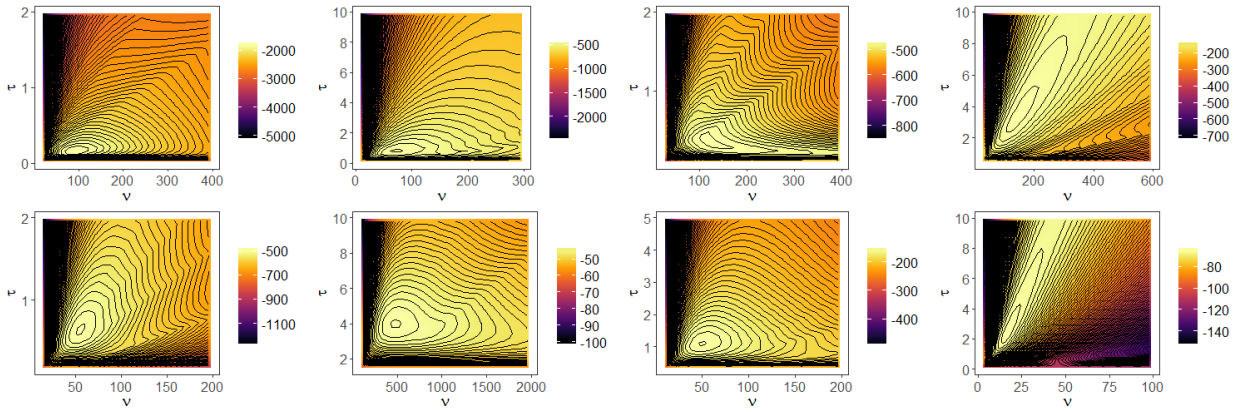


Fig. S2. Likelihood surfaces. Demographic inference with the synonymous SFS for the entire Old World collection and each of the seven lineages was performed using methods implemented in ∂a∂i. Heatmaps of $\log_{10}$ likelihood values (see scale bars) over a range of values for two demographic parameters in a model of instantaneous expansion are plotted: generations since expansion (τ) and N_e/N_{anc} (ν). From left to right, top to bottom: Old World collection (all isolates), L1, L2, L3, L4, L5, L6, and L7.

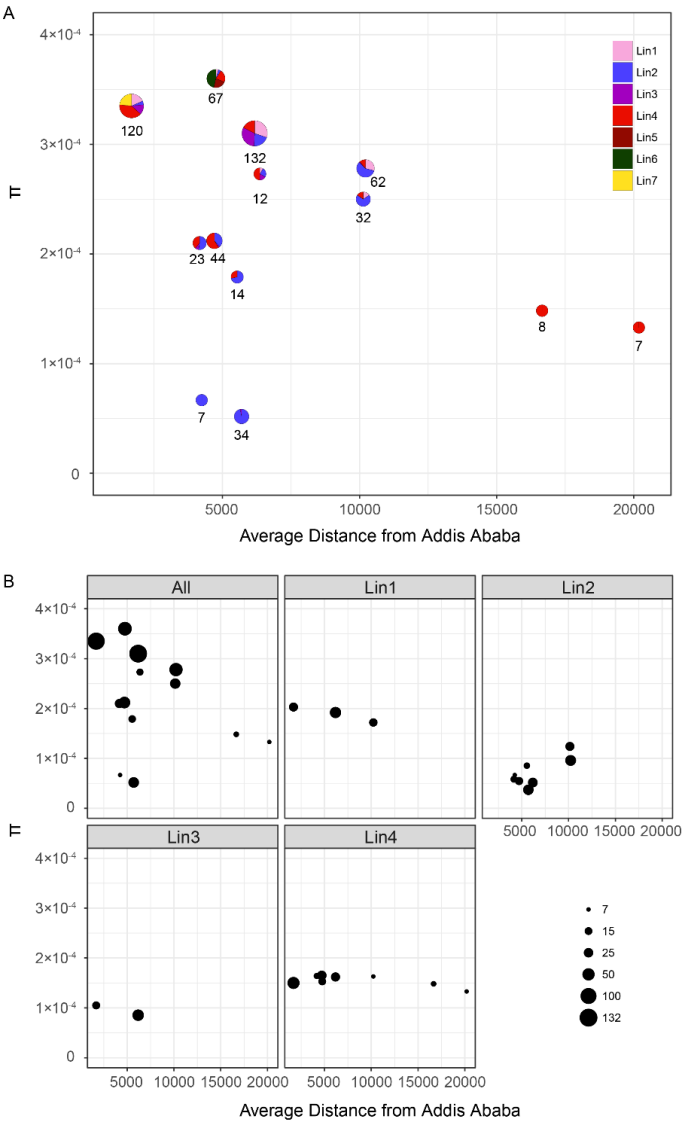
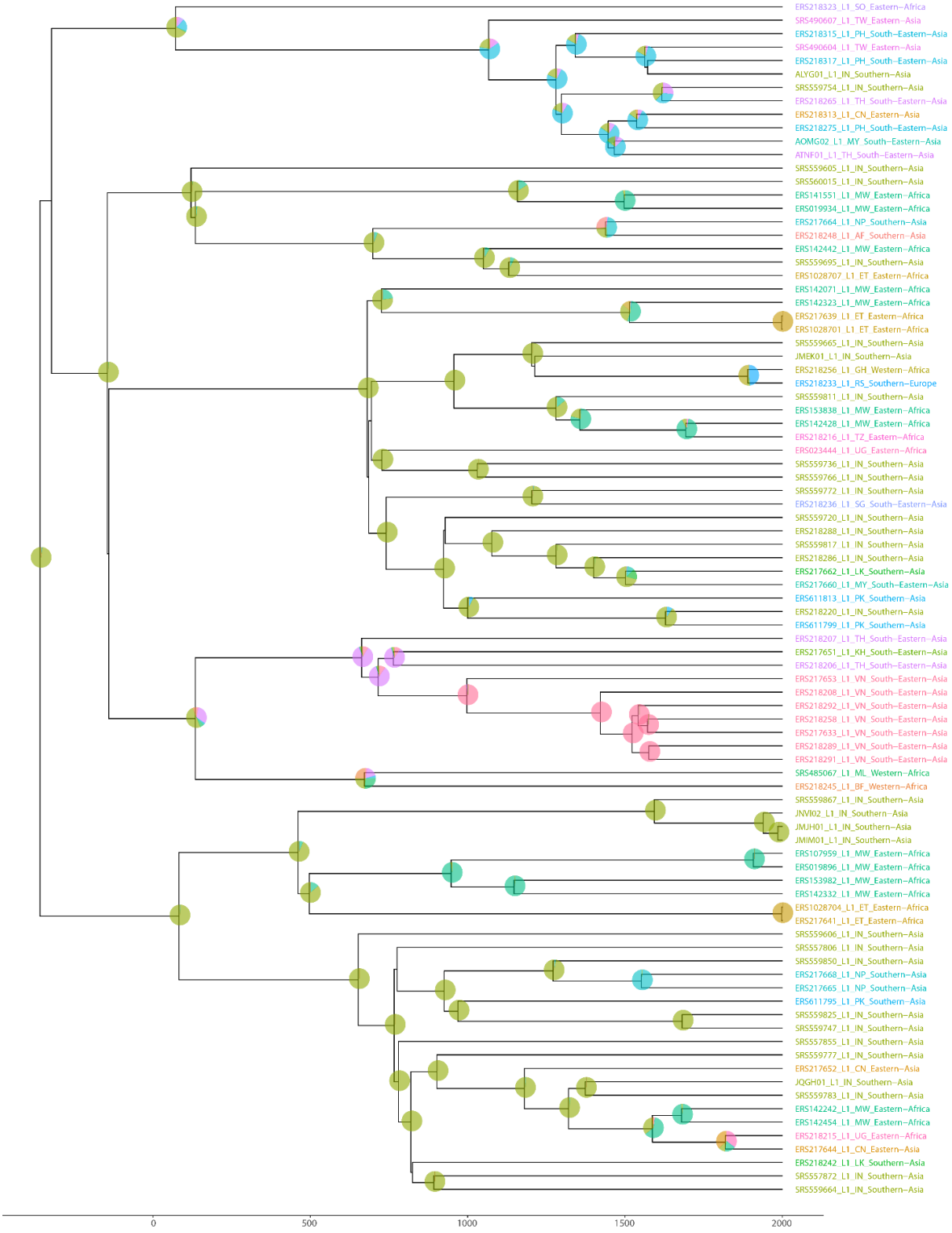


Fig. S3. Relationship between genetic diversity and geographic distance from Addis Ababa (A) Diversity as a function of distance for all isolates among 13 UN subregions. Treating each UN subregion as a population, nucleotide diversity (π) was compared to the mean distance of isolates from the region to Addis Ababa (see *Methods*). Point size reflects the number of isolates per subregion and colors reflect the relative proportions of each lineage (see key). The number of isolates per subregion are denoted near points. (B) Diversity as a function of distance for all isolates and isolates belonging to particular lineages. Individual lineages in each UN subregion were treated as a population and nucleotide diversity (π) was compared to the mean distance of said isolates from the region to Addis Ababa (see *Methods*); population groupings resulting in less than seven isolates were not included. Point size reflects the number of isolates (see key). Data pointsare in Dataset S1.

924

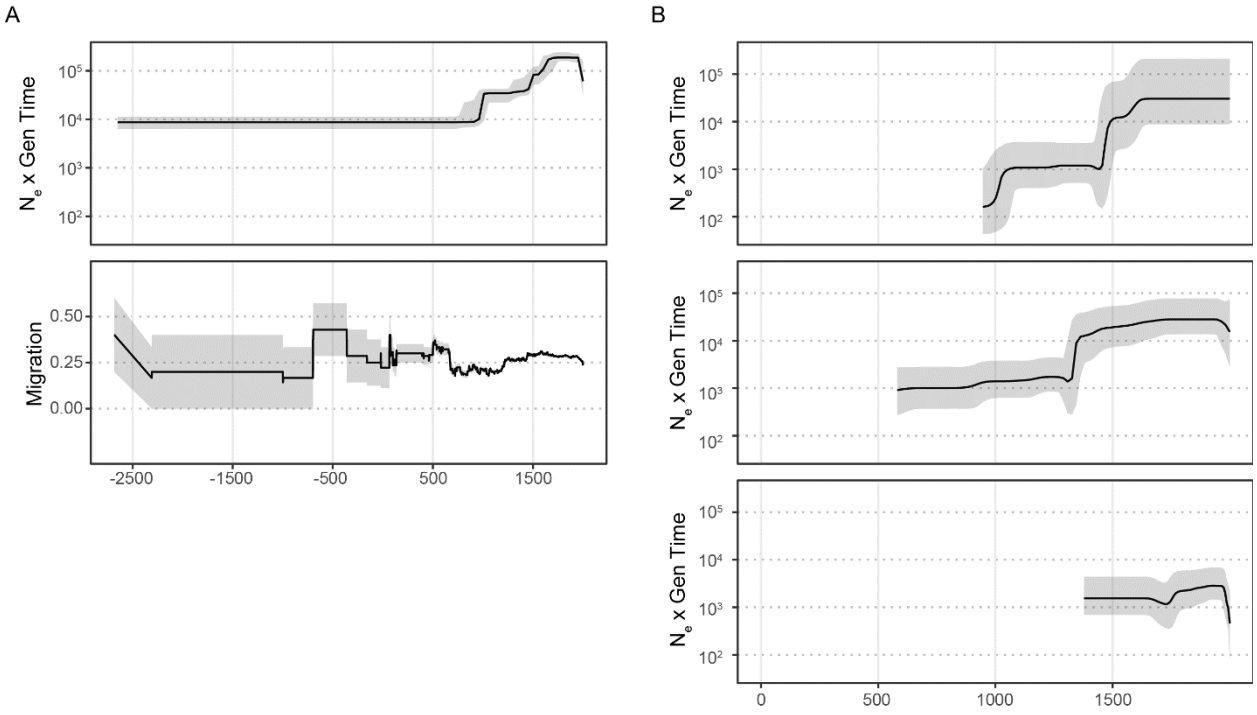


925
926

927 **Fig. S5.** Maximum clade credibility tree of lineage 1 *M.tb* genomes. Pie charts at nodes
928 are colored according to the location state probabilities (country). Tip labels are colored
929 according to the country of origin where the isolate was obtained.

Fig. S6. Maximum clade credibility tree of lineage 4 *M.tb* genomes. Pie charts at nodes are colored according to the location state probabilities (country). Tip labels are colored according to the country of origin where the isolate was obtained.

941



942
943

944 **Fig. S7.** Demographic histories of *M.tb*. (*A*) Old World collection. Top panel – Bayesian
945 skyline plot (BSP) shows the inferred change in effective population through time. Black
946 lines denote median N_e and gray shading the 95% highest posterior density. Bottom
947 panel - migration rate through time inferred from the phylogeographic analysis (see
948 *Methods*). Grey shading depicts the rates inferred after the addition or subtraction of a
949 single migration event, demonstrating the uncertainty of rate estimates from the early
950 history of the phylogeny. (*B*) BSPs for Lineages 5-7 (top to bottom, respectively).
951 Lineages 5-7 are only found in only one subregion each throughout their phylogenies
952 resulting in a migration rate of zero through time. Dates are shown in calendar years
953 and are based on scaling the phylogeny with a substitution rate of 5×10^{-8}
954 substitutions/site/year.

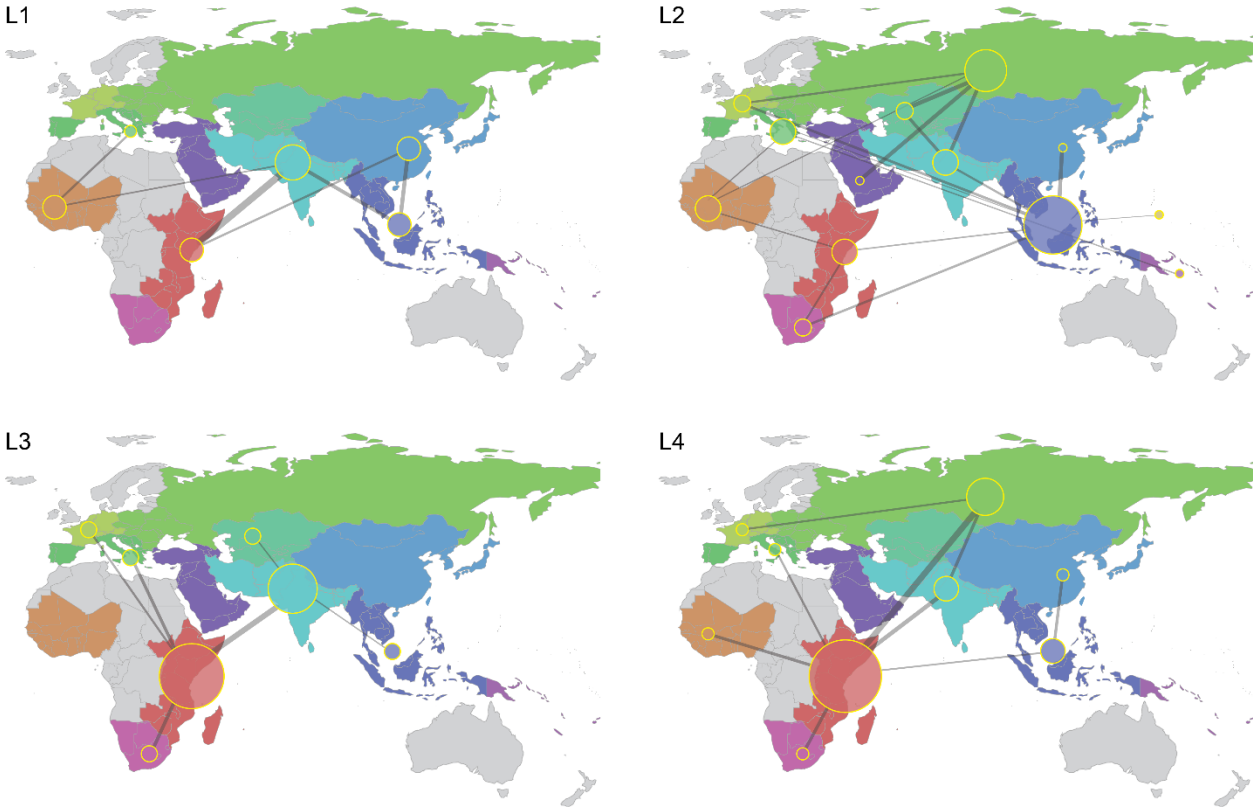


Fig. S8. Relative rates of migration between UN subregions. Relative median rates of migration between geographic regions sampled every 10,000 states in the Bayesian analysis for individual lineages 1-4 are displayed by the thickness of the line connecting the regions. Node size reflects the relative degree of connectivity within each analysis (how many connections the region shares).

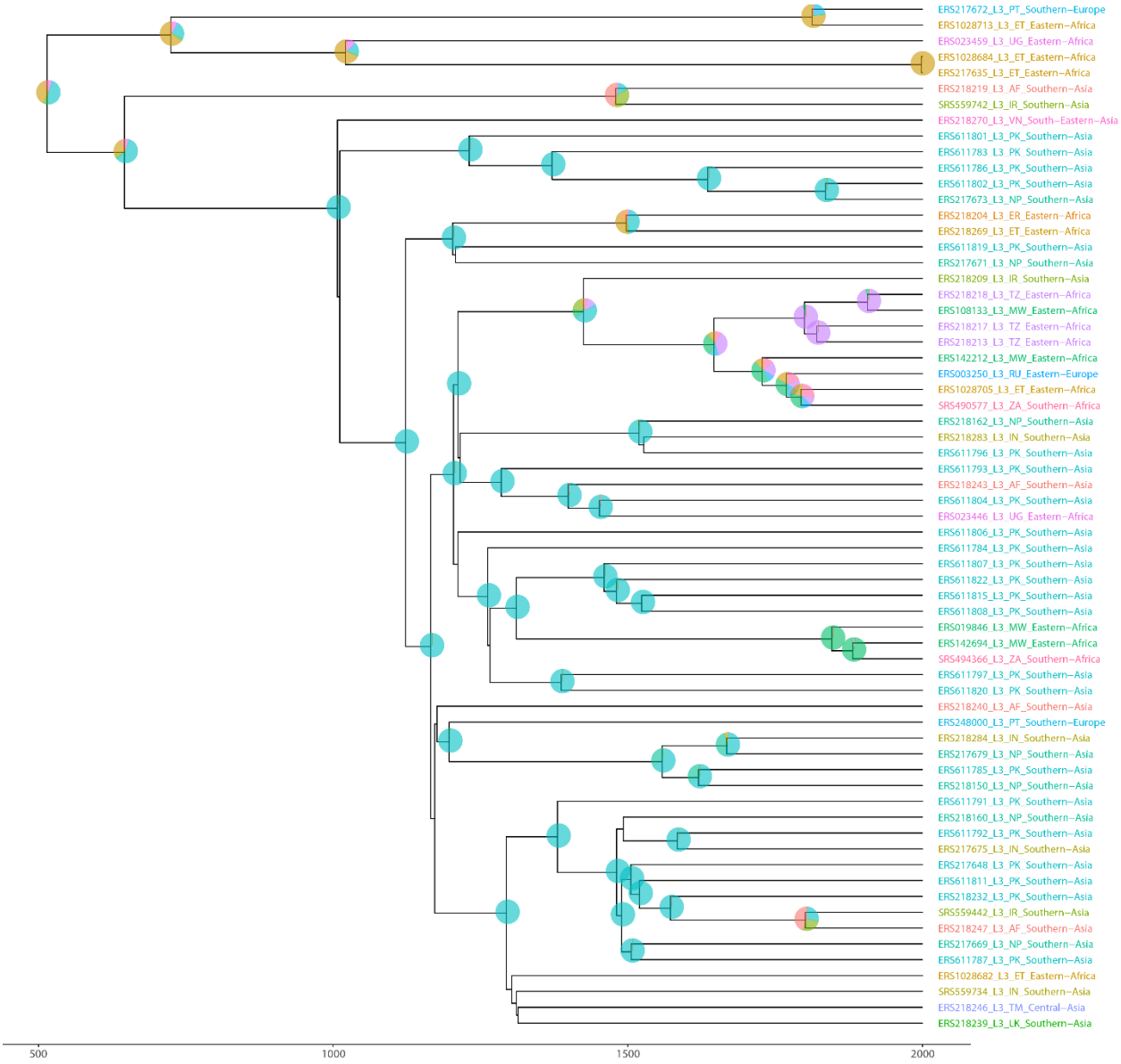


Fig. S9. Maximum clade credibility tree of lineage 3 *M.tb* genomes. Pie charts at nodes are colored according to the location state probabilities (country). Tip labels are colored according to the country of origin where the isolate was obtained.

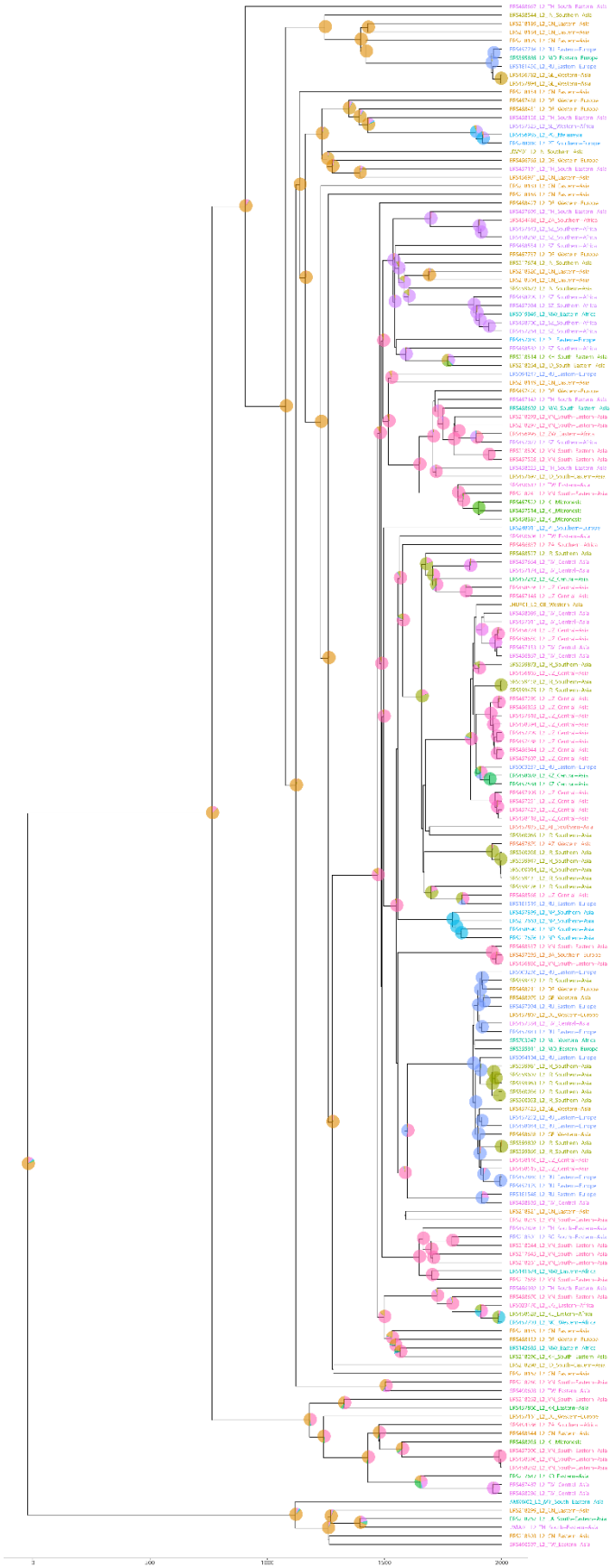


Fig. S10. Maximum clade credibility tree of lineage 2 *M.tb* genomes. Pie charts at nodes are colored according to the location state probabilities (country). Tip labels are colored according to the country of origin where the isolate was obtained.

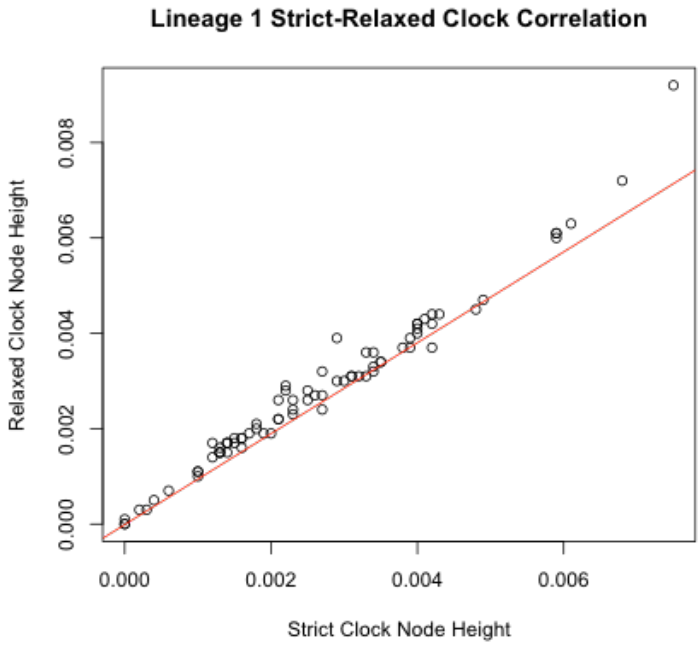


Fig. S11. Correlation of Lineage 1 node heights estimated under strict and relaxed (uncorrelated lognormal clock). R^2 for the correlation is 0.9534. Analyses were performed in BEAST using the same conditions as described in the main text.

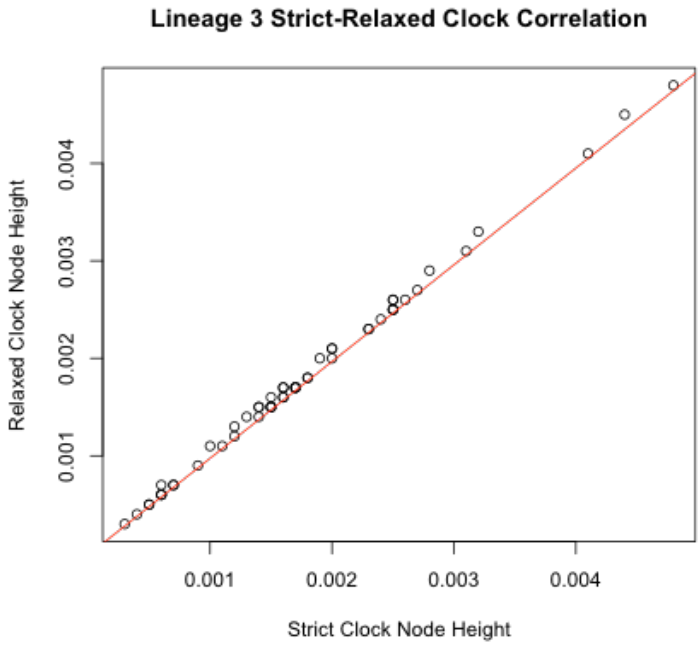


Fig. S12. Correlation of Lineage 3 node heights estimated under strict and relaxed (uncorrelated lognormal clock). R^2 for the correlation is 0.9925. Analyses were performed in BEAST using the same conditions as described in the main text.

Table S2. Summary of BEAST analyses presented. All analyses were performed using the general time reversible model of nucleotide substitution with a gamma distribution to account for rate heterogeneity between sites and a strict molecular clock. Demographic model, discrete traits, and whether the Bayesian stochastic search variable selection method (BSSVS) was implemented are listed in the table as well as the results derived from the respective analyses.

Dataset	Demographic Model	Discrete Trait	Molecular Clock	BSSVS	Results derived
Old World collection	BSP	UN subregion	strict	Yes	TMRCAs, phylogeography (UN subregion), migration through time, Relative rates of migration between UN subregions
Individual lineages (1-7)	BSP	UN subregion	strict	No	Tree structures, BSPs
Individual lineages (1&3)	BSP	UN subregion	relaxed	No	Tree structures, BSPs
Individual lineages (1-4)	constant	ISO2 code (country)	strict	No	Phylogeography (country)
Individual lineages (1-4)	BSP	UN subregion	strict	Yes	Relative rates of migration between UN subregions

Other Supporting Information Files:

[Dataset S1.txt]

Dataset S1. Data used to assess genetic diversity as a function of geographic distance from Addis Ababa.

[DatasetS2_wAmericas.xlsx]

Dataset S2. Old World and Americas *M.tb* collections and associated meta information.

[FigS4_OldWorld_UN.pdf]

Fig S4. Maximum clade credibility tree of 552 *M.tb* genomes. Bayesian analysis was performed using the general time reversible model of nucleotide substitution under the Gamma model of rate heterogeneity, a strict molecular clock, and a Bayesian skyline plot (BSP) demographic model. Time scale is based on a substitution rate of 5×10^{-8} substitutions/site/year. Pie charts at nodes are colored according to the location state probabilities based on the UN geoscheme. Tip labels are colored according to the subregion of the isolate origin.

[TableS1.xlsx]

Table S1. Variable sites in loci associated with drug resistance.