HEALTH AND MEDICINE

Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation

Ola B. Brynildsrud[1], Caitlin S. Pepperell[2,3], Philip Suffys[4], Louis Grandjean[5], Johana Monteserin[6,7], Nadia Debech[1], Jon Bohlin[1], Kristian Alfsnes[1], John O.-H. Pettersson[1,8,9,10], Ingerid Kirkeleite[1], Fatima Fandinho[11], Marcia Aparecida da Silva[11], Joao Perdigao[12], Isabel Portugal[12], Miguel Viveiros[13], Taane Clark[14,15], Maxine Caws[16,17], Sarah Dunstan[18], Phan Vuong Khac Thai[19], Beatriz Lopez[6], Viviana Ritacco[6,7], Andrew Kitchen[20], Tyler S. Brown[21], Dick van Soolingen[22], Mary B. O'Neill[3,23]*, Kathryn E. Holt[14,24], Edward J. Feil[25], Barun Mathema[26], Francois Balloux[27], Vegard Eldholm[1†]

On the basis of population genomic and phylogeographic analyses of 1669 *Mycobacterium tuberculosis* lineage 4 (L4) genomes, we find that dispersal of L4 has been completely dominated by historical migrations out of Europe. We demonstrate an intimate temporal relationship between European colonial expansion into Africa and the Americas and the spread of L4 tuberculosis (TB). Markedly, in the age of antibiotics, mutations conferring antimicrobial resistance overwhelmingly emerged locally (at the level of nations), with minimal cross-border transmission of resistance. The latter finding was found to reflect the relatively recent emergence of these mutations, as a similar degree of local restriction was observed for susceptible variants emerging on comparable time scales. The restricted international transmission of drug-resistant TB suggests that containment efforts at the level of individual countries could be successful.

INTRODUCTION

Tuberculosis (TB) takes more lives than any other infectious disease. Global TB burden has declined slowly over the past decade, but the rise of antimicrobial resistance (AMR) constitutes a significant obstacle to TB control in the absence of an effective vaccine. In recent years, a number of attempts have been made to reconstruct the evolutionary history of *Mycobacterium tuberculosis* (*M.tb*) and its association with humans. One genome-based study hypothesized that *M.tb* spread out of Africa together with early humans (*1*), whereas a later study, using ancient DNA samples for temporal calibration, suggested a far younger most recent common ancestor (MRCA) of extant *M.tb* 4000 to 6000 years ago (*2*). Among seven recognized *M.tb* lineages, lineage 4 (L4) is the most widely dispersed, affecting humans across the world. Here, relying on a collection of 1669 L4 genomes, including hundreds of novel genomes from the Americas, we set out to reconstruct the migration history of L4 and assess the impact of migration on the spread of AMR. We find that repeated sourcing from Europe has been the main driving force for the global expansion of L4, with intense dispersal to Africa and the Americas concomitant with European colonizing efforts ca. 1600–1900 CE. We also find that the rise of multidrug-resistant TB (MDR-TB) in recent decades is overwhelmingly a local phenomenon, in the sense that resistant clones have emerged repeatedly in a wide variety of locations, while migration of resistant strains seems to have played a marginal role in shaping the observed L4 AMR landscape.

RESULTS AND DISCUSSION

Global diversity of L4

In total, 1669 genomes representing clinical *M.tb* isolates from 15 countries were included in the study (Fig. 1). After down-sampling

[1]Division of Infectious Diseases and Environmental Health, Norwegian Institute of Public Health, Lovisenberggata 8, 0456 Oslo, Norway. [2]Division of Infectious Disease, Department of Medicine, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53726, USA. [3]Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53726, USA. [4]Laboratory of Molecular Biology Applied to Mycobacteria, Oswaldo Cruz Institute, Avenida Brasil 4365, C.P. 926, Manguinhos 21040-360, Rio de Janeiro, Brazil. [5]Department of Paediatric Infectious Diseases, Imperial College London, W2 1NY, London, UK. [6]Instituto Nacional de Enfermedades Infecciosas, ANLIS Carlos Malbran, Buenos Aires, Argentina. [7]Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Buenos Aires, Argentina. [8]Department of Medical Biochemistry and Microbiology, Zoonosis Science Center, Uppsala University, Uppsala, Sweden. [9]Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Life and Environmental Sciences and Sydney Medical School, The University of Sydney, Sydney, New South Wales 2006, Australia. [10]Public Health Agency of Sweden, Nobels vg 18, SE-171 82 Solna, Sweden. [11]Laboratorio de Bacteriologia da Tuberculose, Centro de Referłncia Professor Helio Fraga-Jacarepagu, Estrada de Curicica 2000, Brazil. [12]Instituto de Investigao do Medicamento, Faculdade de Farmcia, Universidade de Lisboa, Lisboa, Portugal. [13]Unidade de Microbiologia Medica, Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisboa, Portugal. [14]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK. [15]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK. [16]Liverpool School of Tropical medicine, Department of Clinical Sciences, Liverpool, UK. [17]Birat-Nepal Medical Trust, Lazimpat, Kathmandu, Nepal. [18]Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia. [19]Pham Ngoc Thach Hospital for TB and Lung Diseases, Ho Chi Minh City, Vietnam. [20]Department of Anthropology, University of Iowa, Iowa City, IA 52242, USA. [21]Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA 02114, USA. [22]Center for Infectious Disease Research, Diagnostics and Perinatal Screening, National Institute for Public Health and the Environment, P.O. Box 1, 3720 BA Bilthoven, Netherlands. [23]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA. [24]Department of Biochemistry and Molecular Biology and Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia. [25]Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK. [26]Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, USA. [27]UCL Genetics Institute, University College London, London WC1E 6BT, UK.
*Present address: Unit of Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France.
†Corresponding author. Email: elve@fhi.no

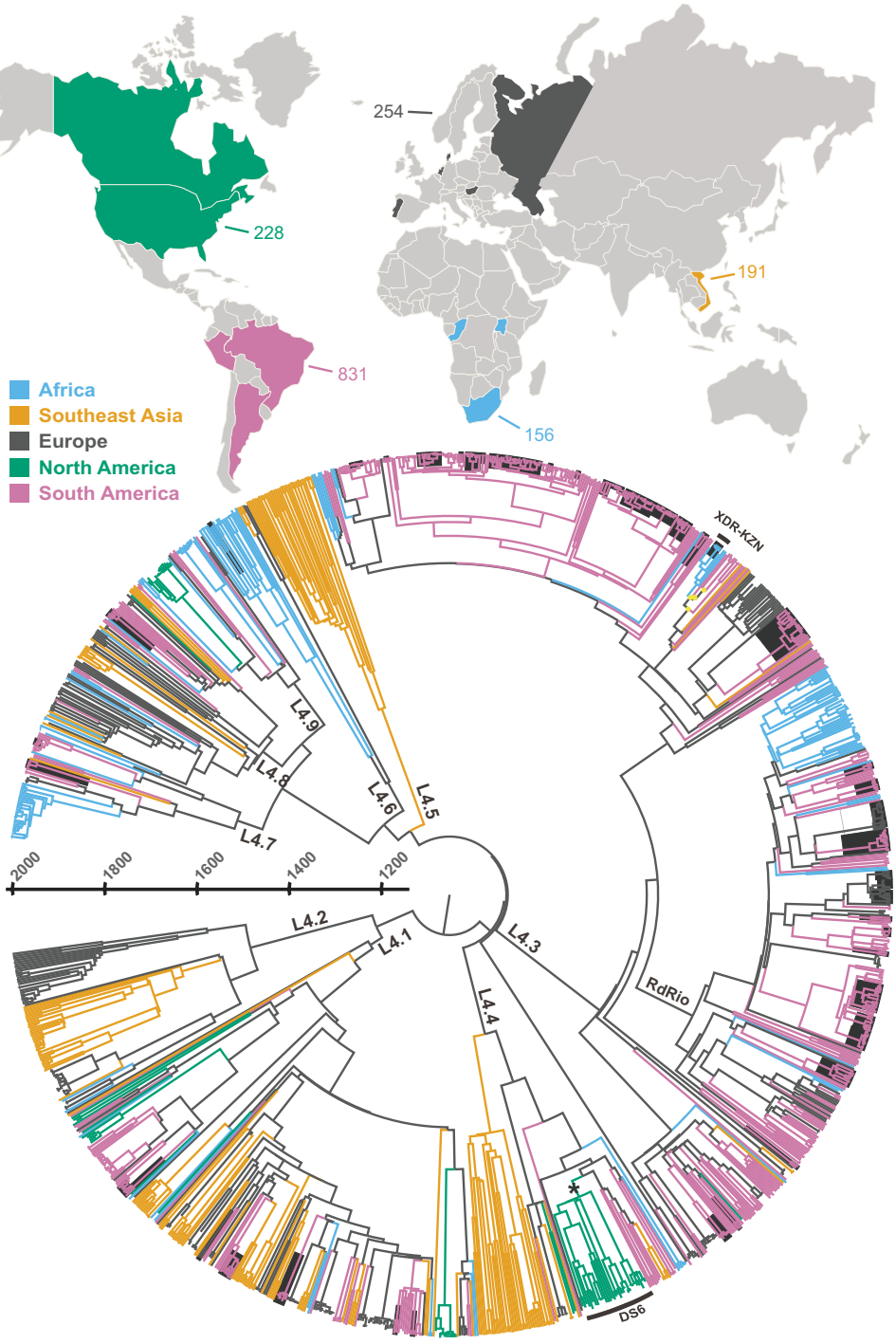

Fig. 1. Sampling overview and phylogeography of the global L4 dataset. In the map, sampled countries are colored by continent, and sample sizes are indicated. In the temporal phylogeny, branches are colored to match the most likely geographic location inferred using BASTA. MDR clusters identified in the dataset are highlighted with black background shading. A large black asterisk highlights the branch leading to the $DS6_{Quebec}$ clade that was used to assess robustness of dating analyses, whereas yellow dots indicate independent introductions of the KZN ancestor to South Africa and South America (see main text for details).

of densely sampled outbreaks (see Materials and Methods), 1205 isolates remained. To get an overview of global patterns of genomic diversity and strain distribution, we assigned each genome to a sublineage based on the Coll scheme (*3*) and mapped the sublineage annotations on the temporal phylogeny (Fig. 1). In general, sublineages were found to be widely dispersed, but clear patterns of geographic structure are discernible, as noted by Stucki *et al.* (*4*): L4.5 was restricted to Southeast Asia (Vietnam), whereas L4.3 (also termed LAM) was underrepresented in the country (Fig. 2). We also observed early, distinct splits within sublineages 4.2 and 4.4;

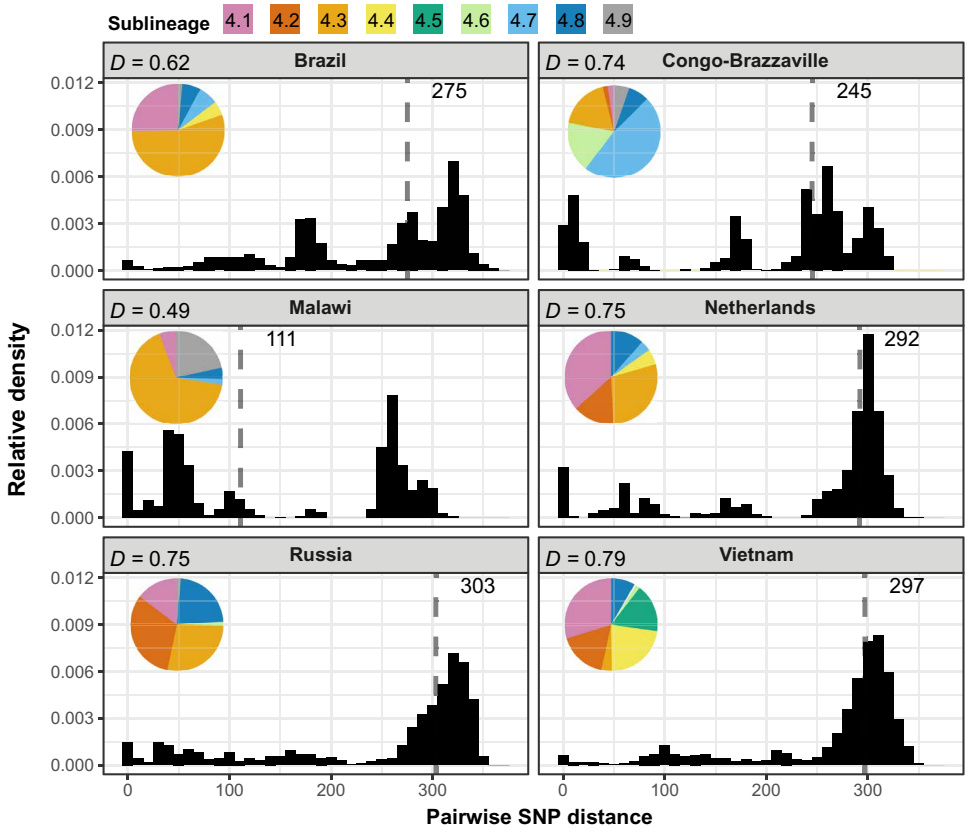

Fig. 2. Within-country diversity as assessed by mean pairwise SNP distances (only including well-sampled countries). Vertical dashed lines indicate median values. Embedded pie charts summarize sublineage distribution within each country. The Simpsons diversity index (1-D) was calculated at the level of subspecies and the estimate indicated in the top of each county panel (where higher values correspond to increased diversity).

L4.2 consists of a Vietnamese cluster nested within the otherwise strictly European sublineage, whereas an early branching event separates L4.4 into two clades, one exclusively detected in Vietnam and the other global in its distribution.

To assess the overall L4 diversity as a function of geography, we investigated the distributions of pairwise single-nucleotide polymorphism (SNP) differences in countries where sampling was deemed to be sufficiently dense and representative (Fig. 2). Russia, the Netherlands, and Vietnam were characterized by the highest diversity in circulating L4 strains, both in terms of median pairwise distance and Simpson's diversity (which also takes sublineage distribution into account). Comparing the patterns of diversity in Malawi and Vietnam, representing large collections from Karonga district and Ho Chi Minh City, respectively (*5*, *6*), the SNP distance distributions indicate more ongoing transmission of L4 in Malawi (where pronounced peaks in the lower tail of the pairwise distance distribution is visible) relative to Vietnam.

Phylogeographic inference

Next, we performed analyses to formally assess the phylogeographic history of L4. We used both discrete trait analyses (DTA) (*7*) and Bayesian-structured coalescent approximation (BASTA) (*8*). For both analyses, a temporal phylogeny inferred with Bayesian evolutionary analysis sampling trees (BEAST) v1.8.4 (*9*) was used as an input tree, and a sampling location was used to assign samples to one of five regions: Europe, Africa, Southeast Asia, South America, and North America [Russian isolates were assigned to Europe, since all isolates were sampled West of Ural, in accordance with United Nation's (UN) geoschemes]. The L4 MRCA was estimated to have existed around 1096 CE [95% highest posterior density (HPD), 955 to 1231] (see the Supplementary Materials). Applying DTA, the geographic location of the MRCA was estimated to be Europe with high confidence (posterior probability = 0.92).

BASTA initially placed the root location in South America, which in light of the estimated age of L4 is highly unlikely, and also necessitated migration events to the Old World in the centuries immediately preceding the Columbian discovery of the Americas. We thus performed individual analyses with the root location forced to be any of the three Old World continents to which our samples belonged. In these analyses, a European root location was overwhelmingly favored (see Materials and Methods). The resulting migration matrices estimated by DTA and BASTA were largely concordant (fig. S5), both methods inferring Europe as playing a pivotal role in the global dispersal of L4. The central role of Europe in the global expansion of L4 was most pronounced in the BASTA inferences, where out-of-Europe migration was found to be almost singlehandedly responsible for the current geographic range of L4 (Fig. 3A). However, "Europe" should not be interpreted in the strict sense and probably captures interactions in the Middle ages with western Asia and North Africa as well. Thus, our study does not contradict an origin around the Mediterranean, as suggested in a recent study (*10*).

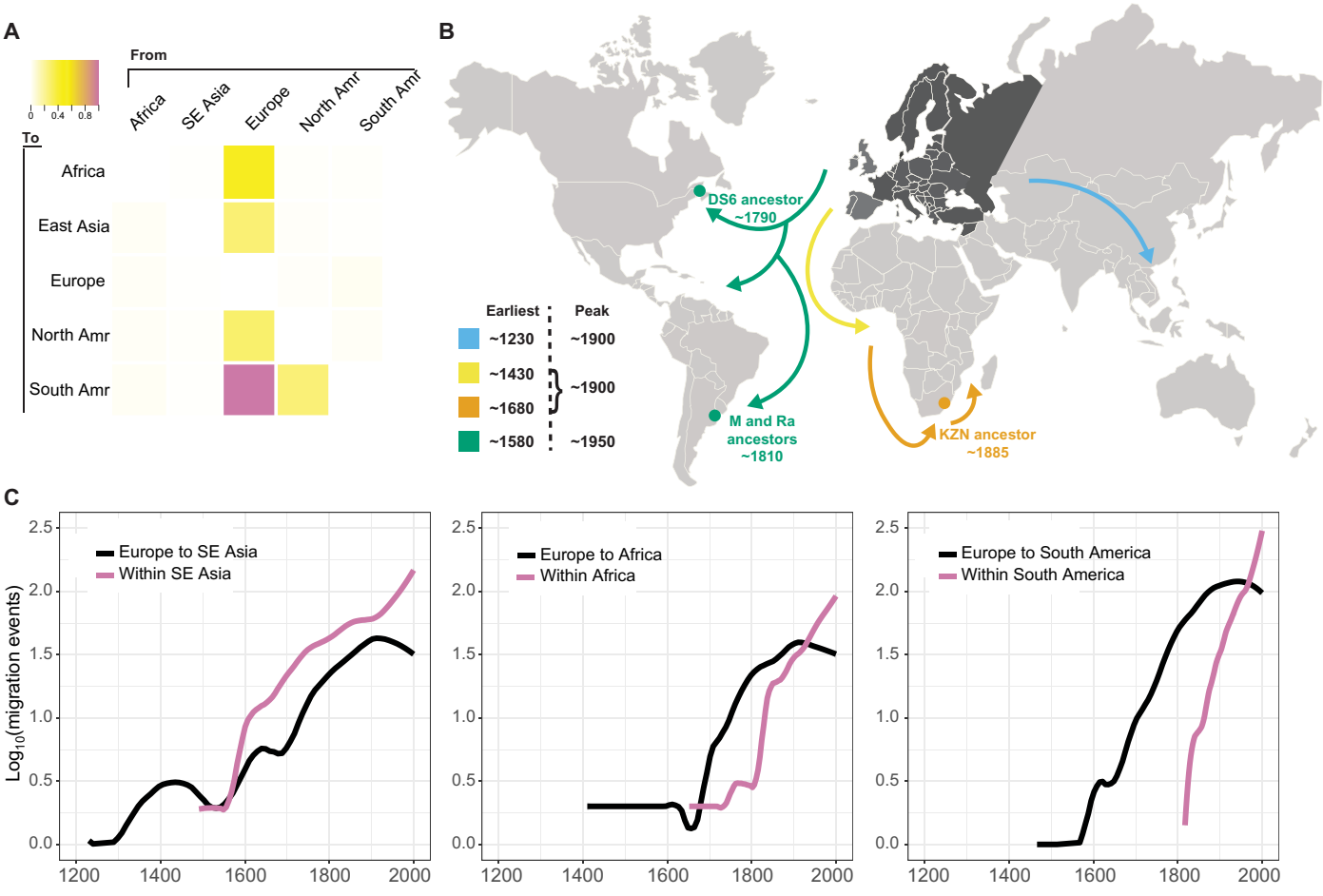

Fig. 3. L4 migration. (**A**) Heatmap summarizing the overall migration load between continents as inferred in BASTA. (**B**) Temporal overview of L4 migration out of Europe. The establishment of strains of interest discussed in the text is highlighted. As the exact timing of the first American migrations was uncertain, the mean of the first three inferred migration events to each of the two subcontinents is reported as an approximation of the earliest migration events to the Americas. (**C**) Out-of-Europe migration to Southeast Asia, Africa, and South America over time. The plots also show within-continent migration/transmission in the receiving continents to illustrate the relative importance of repeated L4 import on continental L4 load over time. SE, southeast.

Patterns of L4 migration through time

To investigate the migration history of L4 in a temporal context, we analyzed the inferred load and direction of migration over time (see Materials and Methods). As it was clear that out-of-Europe migrations were the overwhelming driver of global L4 range expansion (Fig. 3A), we illustrate migration from Europe to the other continents through time in Fig. 3B. In parallel, we analyzed migration events within the receiving continents to get a picture of the relative importance of import versus intracontinental transmission (Fig. 3C).

Our phylogeographic analyses suggest that the first waves of L4 migration out of Europe were in an eastward direction, with the first migration to Southeast Asia (represented by Vietnam) estimated to the beginning of the 13th century. Here, local populations were quickly established, and internal transmission became dominant by the late 16th century (Fig. 3C). These observations fit well with the observed population structure of the Vietnamese isolates (Fig. 1) and the high diversity observed within the country (Fig. 2). Present-day Vietnam was part of a large French colony termed French Indochina from the late 19th century onward. France was not among the sampled countries in the current study, but focusing on the Netherlands, the most proximal, sampled country, we identified a number of nodes with Vietnamese and Dutch descendants only, the first dated to 356 to 426 years before present (95% HPD), followed by six nodes with 95% HPDs from 114 to 269 years ago. These date ranges fit well with known French-Vietnamese interactions, starting with the arrival of the missionary Alexandre de Rhodes in the 1627 (*11*) followed by French military expansion from the mid 19th century and the formation of French Indochina in 1887.

The next waves of migration were directed toward Africa, with the earliest introduction among the sampled countries inferred in the present-day Republic of Congo (Congo-Brazzaville) in the 15th century and subsequent introductions to South Africa, Uganda, and Malawi (Southern and Eastern Africa) from the late 17th century. The early introduction in Congo is likely a result of the relatively proximity of Congo-Brazzaville to the West African territories, which were first to interact with European explorers. In contrast to what we observed for Southeast Asia, repeated sourcing from Europe seems to have been more important than local

transmission until the 19th century (Fig. 3C). These findings closely mirror the European colonial history in Africa south of the Sahara, with early Portuguese forts and trading posts established on the Gold Coast (present-day Ghana) in 1482, followed by an ever increasing European presence and influence in African coastal regions over the next centuries. Last, this culminated in an all-out land grab termed the "scramble for Africa" in the late 19th century, placing vast portions of the African continent under European control in the form of colonies. Our findings thus suggest an intimate relationship between European colonial expansion and the spread of L4 TB on the African continent. The main form of internal African migration in this time period was connected to the expansion of the Zulu Empire under Shaka (1816–1828), which forced fleeing tribes to migrate northward and eastward. This appears to have been less important in the spread of L4, but internal nodes with descendants in Congo, Uganda, and Malawi dated to the 18th to 20th century point to secondary transmission routes through internal African migration.

The three earliest migration events to the Americas were inferred to have occurred between 1466 and 1593 to South America and between 1566 and 1658 to North America. These migration events occurred along long branches of the phylogeny, so the exact timing cannot be established. The estimates do however suggest that Europeans brought TB to South America relatively soon after the arrival of Europeans on the continent in 1492. An abrupt increase in the flow of L4 into South America is seen from the turn of the 17th century (Fig. 3). Bone pathology and lesions identified in human remains suggest that TB in the Americas might have predated European contact (*12*), but convincing molecular evidence is limited to the identification of *Mycobacterium pinnepedii* infection (a *M.tb* complex member generally restricted to seals and sea lions) in skeletal samples from a Peruvian coastal settlement (*2*). As the original human colonization of the Americas predates the likely age of the *M.tb* MRCA (*2*) by a large margin, the most parsimonious interpretation is that, while animal strains were in circulation in some Native American human populations, *M.tb* sensu stricto was introduced to the Americas by European colonists and followed by continued influx of L4 with subsequent waves of European migrants. The near-complete dominance of L4 in South America (*13*) also supports this notion.

Despite the massive and relentless import of TB, we find that the establishment of detectable local transmission was delayed in South America relative to Africa and Southeast Asia (Fig. 3), which might reflect the massive decline of native populations following European contact. Infectious diseases caused severe die-offs of native populations following the arrival of Europeans (*14*). The toll taken by infectious disease epidemics in Native Americans is ascribed to their vulnerability to a number of pathogens introduced by Europeans and was likely exacerbated by widespread societal collapse and famines. The near-wholesale replacement of Native Americans with Europeans and African slaves in many regions (*14*) might explain the delayed imprints of local TB transmission in our phylogeographic analyses. A figure including migration over time to both North (relatively sparsely sampled after down-sampling of outbreaks) and South America is included in fig. S6.

The phylogeographic analyses also allowed us to shed light on the history of specific strains of interest. In South Africa, the KZN strain is responsible for a devastating epidemic of extensively drug-resistant TB (XDR-TB) (*15*, *16*). We find that the ancestor of the KZN strain was dispersed with Europeans to South Africa about 130 years ago (Fig. 3B). In the same period, about 100 to 150 years ago, there were independent introductions of a closely related strain from Europe to Latin America, providing context to the observation by Lanzas *et al.* (*17*) that strains closely related to the KZN strain are driving an MDR-TB outbreak in Panama.

Another interesting finding concerns the RdRio clade, originally identified as a major cause of TB in Brazil (*18*) but later identified at moderate to high frequencies also in Portugal, the United States, and beyond (*19*, *20*). Our results suggest that the RdRio clade originated in Europe about 350 years ago, followed by multiple introductions to Africa and the Americas from around 250 years ago. To investigate this clade in more detail, we extracted the RdRio subtree from the full L4 phylogeny and performed an independent phylogeographic analysis. This analysis indicated Iberia as a likely origin of RdRio, followed by early expansions to Atlantic South America, Peru, and Southeast Africa (fig. S7).

Furthermore, we find that the ancestors of two major MDR outbreak strains in Argentina, M (*21*) and Ra (*22*), were both introduced to South America approximately 200 years ago. The L4 $DS6_{Quebec}$ clade is common among Aboriginal populations in Ontario, Saskatchewan, and Alberta, as well as French Canadians in Quebec. As substantial contact between these populations was limited to the period of 1710 to 1870 (*23*), this provides a useful interval against which to test our temporal inferences. Our analyses places the MRCA of $DS6_{Quebec}$ in Canada in 1788 [95% HPD, 1739 to 1827] well within this interval. The inferred timing of L4 migration to Africa and the Americas fits remarkably well with the known history of European colonization of the continents. A summary of the direction, intensity, and timing of L4 migration is included in Fig. 3.

Local adaptation and AMR

For clonal, nonrecombining organisms such as *M.tb*, the identification of homoplasic mutations is a powerful way to identify targets of selection (*24*). We identified a total of 733 mutations that had emerged more than once in the L4 dataset. As expected, the top-scoring genes included a number of known AMR genes (dataset S2). We also identified a handful of promoter and nonsynonymous gene mutations (codons 3 and 253) in the lactate dehydrogenase gene *lldD2* that had evolved independently >100 times. A recent study demonstrated that *lldD2* is important for *M.tb* replication within human macrophages by enabling the bacillus to use lactate as a carbon source (*25*). A screen for positive selection in a smaller dataset covering *M.tb* lineages 1 to 6 (*26*) found that the codon 3 mutation had emerged independently in lineages 1, 2, and 4, whereas the codon 253 mutation had emerged repeatedly in L4 and was present in all but a single L2 isolate. In L4, we find that *lldD2* mutations started emerging well before the age of antibiotics (fig. S9) and have emerged across all continents (dataset S3), suggesting parallel local adaptation, most likely to broad changes in host ecology. Last, we assessed the transmissibility of clades with and without *lldD2* mutations, both in terms of number of descendants per node age and in terms of transmission across country borders (see supplementary text). These analyses suggested that there were indeed differences between the groups and that strains harboring *lldD2* promoter mutations carry a significant benefit in terms of transmissibility.

The most serious challenge to TB control efforts is the rise of AMR, which is threatening to reverse the moderate decrease in TB burden obtained over the last decade (*27*). To investigate the role of migration in the spread of AMR, we mapped the time and location of resistance emergence by mapping known resistance mutations on the temporal phylogeny. For clarity, we only included genes relevant for the multi- and extensive drug resistance definitions [MDR-TB, resistance to isoniazid (INH) and rifampicin (RIF); XDR-TB, MDR-TB with additional resistance to fluoroquinolones (FLQ) and one of the injectable drugs kanamycin, amikacin, or capreomycin]. As the major mutations responsible for kanamycin, amikacin, and capreomycin resistance are largely overlapping, we refer to this group as KAN resistance mutations. The stacked bar chart summarizing the inferred timing of individual resistance emergence events (Fig. 4, top) indicates a gradual increase in AMR emergence rate from the 1960s until the late 1990s, after which a plateau is reached. In line with an earlier global study of AMR in *M.tb* (*28*), we find that MDR-TB has emerged repeatedly and independently across geographic regions (Fig. 1).

When extracting the geographic location of isolates descending from resistance nodes (i.e., inherited resistance), we did not identify a single instance of a resistant strain crossing country borders. To understand the underlying cause of this observation, we first investigated whether this reflected a decrease in transmission fitness of resistant strains (see Materials and Methods). This analysis did not indicate any decrease in the transmissibility of resistant strains relative to their susceptible counterparts (Fig. 4, bottom). We thus hypothesized that the failure of resistant strains to cross country borders in our global dataset might simply reflect the young age of these strains. Cross-country and cross-continental migration events as a function of node age and the predicted phenotypic resistance of the strain occupying the node were thus quantified. From Fig. 4 (middle), it is clear that cross-border migration was exceedingly rare among descendants of nodes young enough to be resistance nodes, irrespective of susceptibility profile. Only a single node emerging within the age of antibiotics (post 1945) was found to have descendants that had crossed country borders in our dataset. We thus conclude that the young age of resistance nodes, rather than decreased transmission fitness, explains the lack of observed migration of these strains.

Concluding remarks

As a result of the clonal mode of *M.tb* replication (*29*), efficient adaptive evolution across populations requires the parallel evolution of beneficial mutations. This is indeed the pattern we observe both for adaptive mutations in the lactate metabolism gene *lldD2* and, importantly, the emergence of multidrug resistance across the L4 phylogeny (Fig. 1). AMR emergence within L4 TB reflects local adaptations to near-identical treatment schemes across diverse geographic contexts. There is no doubt that resistant *M.tb* strains can cross country borders and has been observed, e.g., in the case of resistant L2 isolates imported from Eastern to Western Europe (*30*), but we demonstrate that migration has played a negligible role in shaping global AMR patterns in L4. The geographic restriction of resistant strains is indeed striking and suggests that the challenge of AMR can still be tackled efficiently at the level of individual nations. If, however, we fail to act swiftly using the best informed interventions available, then this picture might change rapidly.

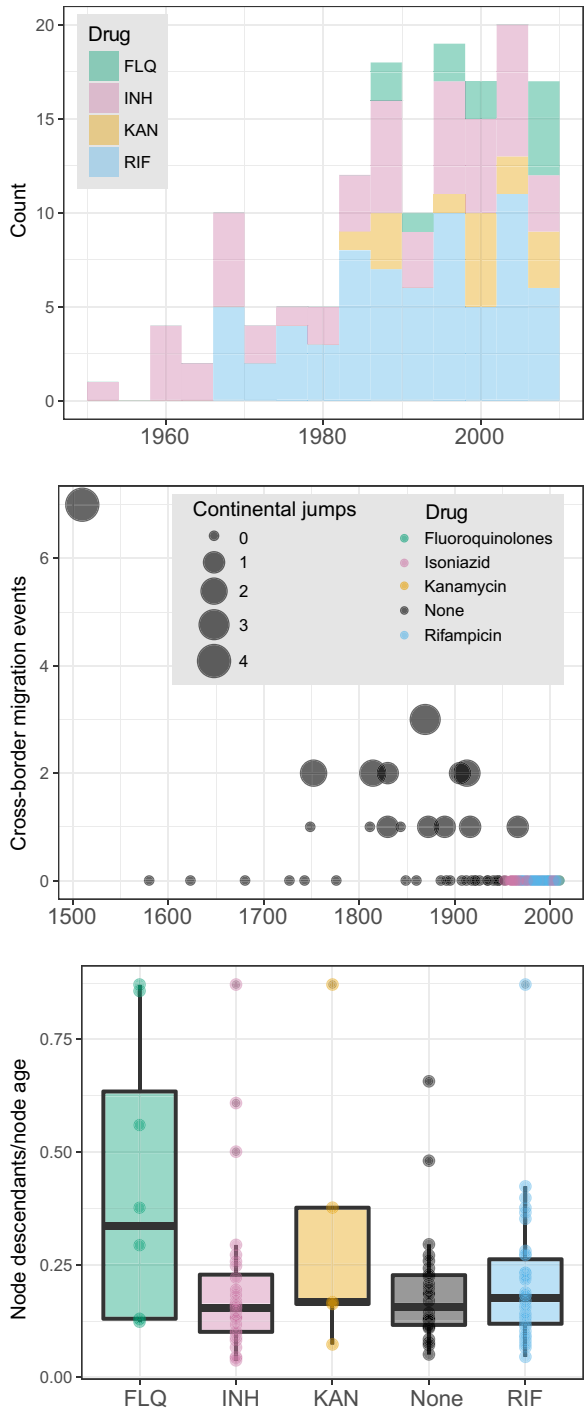

Fig. 4. Transmission of resistance. (**Top**) Independent emergence of AMR over time based on the age of nodes where resistance mutations were inferred to have emerged. (**Middle**) Inferred cross-border transmission of the descendants of susceptible and resistant ancestors as a function of node age. The size of the dots indicates the number of inferred cross-continental migration events occurring among ancestors of each node. Individual dots are colored by the drug to which they cause resistance. (**Bottom**) The number of descendants divided by node age for inferred nodes with or without resistance mutations as a proxy for transmissibility.

MATERIALS AND METHODS

Sample collection

To aid population genomic and phylogeographic inferences, we aimed to include large and representative sample collections covering the widest obtainable temporal and geographic range for the study. We included L4 genomes from recently published studies from Argentina (*21*), Canada (*31*), Congo-Brazzaville (*32*), Malawi (*33*), Netherlands (*34*), Portugal (*35*), Russia (*30*), South Africa (*15*), Uganda (*28*), United Kingdom (*30*), United States (*36*), and Vietnam (*6*). To further improve temporal structure and resolution, we also included three genomes isolated from 18th century Hungarian mummies (*37*) and genomes from Denmark sampled in the 1960s and 1990s (*38*). In addition to published genomes, we included 627 previously unpublished genomes from the Americas (Brazil, Peru, Argentina, United States, and Canada). Sequencing libraries were prepared as described previously (*39*) and sequenced on an Illumina platform. Inclusion criteria for individual genomes were as follows: (i) The genome was annotated with a minimum of metadata (sampling year and country of origin), or this information could be obtained from the published articles or by correspondence with personnel involved in the sequencing efforts. (ii) The genome had to be L4, as verified by the program TB profiler (*40*). Genomes that were determined to be of mixed origin were also excluded. (iii) If a genome was determined to be a duplicate of an already included isolate, then it was not included. (iv) All of the following quality control criteria had to be met: (iv-a) When mapping reads against a H37Rv reference genome, a minimum genome coverage of 90% had to be reached. (iv-b) The average read depth across the reference had to exceed 20. We allowed sequence data from different Illumina platforms and with different read lengths. A total of 1669 genomes passed all inclusion criteria (dataset S1).

Samples from Argentina and from a study of an Inuit community in Quebec, Canada (*31*) were mainly from outbreaks and thus harbored limited within-population diversity. To control for the effect of densely sampled genomes on certain analyses, we additionally created a down-sampled genome collection, where we allowed only five isolates from each of the Argentinian outbreaks (M and Ra) and 10 from the outbreak in Quebec. These genomes were randomly sampled from the full collection by a random number generator. In total, the down-sampled isolate contained 1207 isolates, including the H37Rv reference and an L3 outgroup isolate.

Variant calling

The snippy pipeline v3.1 (https://github.com/tseemann/snippy) was used for variant calling. Briefly, this entailed mapping reads against H37Rv with the BWA mem algorithm (v.0.7.15-r1110) and marking split hits as secondary and then calling variants with SAMtools v1.3 and including only reads with a mapping quality of 60 or higher. Variants were then filtered further using FreeBayes (v1.0.2) with a ploidy of 1 and options to exclude (i) alleles if a supporting base quality is less than 20 or the coverage less than 10, (ii) alignments if the mapping quality is less than 60, and (iii) alleles if the fraction of reads in support of a SNP is not at least 90%. The binomial priors about observation expectations were turned off. The program snpEff v4.1l was then used to annotate SNPs, turning off downstream, upstream, intergenic, and 5′ untranslated region (5′UTR) and 3′UTR changes. A whole-genome alignment of all genomes was then built using snippy core, with a minimum coverage depth of 10 to consider a region part of the core.

The genome sequences from the Hungarian mummies were resolved in a different manner: Sequence read archives from study PRJEB7454 (*37*) were mapped to H37Rv (*41*). A minimum read length of 35 bp (base pairs) and a minimum mapping quality of 30 were imposed. Pilon was run with the following parameters: – variant – mindepth (10) – minmq (30) – minqual (30), and the number of reads found supporting each allele across all variant sites were manually inspected. Three genotypes were inferred from two high-coverage sequencing runs [ERR651000 (individual 68) and ERR651004 (individual 92)]. Genotypes were distinguished on the basis of allele frequencies. For individual 68, genotype 1 was deduced by alleles found between 55 and 65% and genotype 2 by alleles found between 35 and 45%. Variants at frequencies of ≥95% were considered fixed between the mixed infecting strains. Variants segregating at other frequencies were treated as ambiguous/missing data. For individual 92, we also found evidence indicative of a mixed infection; however, we only felt confident with the most called genotype, which comprised >90% of reads. All variants at ≥95% were called. All other sequencing runs from the project were of low coverage and excluded from the analyses.

We then used an in-house python script (available at https://github.com/admiralenola/globall4scripts) to exclude, from the alignment, SNPs matching any of the following criteria: (i) located in a known repetitive region (e.g., PE/PPE genes, annotation file available at github repository), (ii) the proportion of ambiguous calls at the locus exceeded 1%, and (iii) the position was no longer polymorphic after pruning of outbreak isolates (only applied to the down-sampled dataset). The final alignment of SNPs consisted of 22,912 sites, with 9313 of these being parsimony informative.

Initial phylogenetic analysis

The program ModelFinder (*42*) as implemented in IQ-TREE was used to infer the optimal substitution model for our genome alignment. A maximum-likelihood phylogenetic tree was then built using the program IQ-TREE v1.4.3 (*43*), applying the generalized time-reversible (GTR) model with four gamma categories for rate variation, ascertainment bias on, and 1000 ultrafast bootstrap replicates (*44*). The L3 isolate SRR1188186 (Quebec, Canada) was used as an outgroup.

AMR gene mutations

Many *M.tb* AMR mutations are known, but the contribution to the resistance phenotype and penetrance remains unclear for a substantial fraction of variants. We therefore included only high-likelihood AMR mutations relevant for the MDR and XDR definitions. All loci were selected a priori for their perceived relevance and strength of association to AMR. For INH, we included the *katG* S315T mutation and any nonsense or frameshift mutations in the gene plus the classical −15 and −8 *inhA* promoter mutations. For RIF, we included all nonsynonymous mutations in the resistance-determining region (amino acids 426 to 450) of *rpoB*. For kanamycin/amikacin/capreomycin, we included *eis* promoter mutations in positions −14, −12, and −10 (*35*, *45*), as well as mutation in position 1401 of *rrs*. For FLQ, we restricted the analysis to a manually curated collection of high-likelihood mutations, namely, *gyrB* mutations leading to amino acid substitution at positions 461, 499 to 501, and 642, as well as *gyrA* mutations resulting in amino acid substitutions at positions 88 to 94 and 288 (*46*, *47*).

Resistance-associated loci were extracted from the whole-genome alignment using EMBOSS v6.6.0.0 (*48*) using their H37Rv coordinates. The loci were translated to protein, and sequences were sorted by alleles. Mutations were manually annotated on the phylogenetic tree using simple parsimony (e.g., an internal node was inferred to have allele A if all descendent nodes had allele A.). The figure was drawn using iTOL (*49*).

Population genomic inferences

Gene-wise estimates of nucleotide diversity (π), Watterson's theta (θ), and Tajima's D were computed for each gene by parsing the whole-genome alignment into genes using EMBOSS (*48*) and using the python package EggLib (*50*) on individual gene alignments. F_{st} values were also calculated using EggLib, specifying groups by (i) country and (ii) sublineage as identified by the TB profiler. Within each country, we calculated genomic pairwise Hamming distances using the program snp-dists (https://github.com/tseemann/snp-dists).

For each gene, we additionally calculated the number of homoplasies and number of homoplasic sites using a Fitch downpass algorithm, as implemented in Dendropy (*51*). Gene-wise parsimony scores were calculated by identifying homoplasic mutations in each gene and summing the Fitch parsimony scores (i.e., the minimum number of independent emergences of each mutation). To investigate potential alterations of fitness induced by the various mutations in and immediately upstream of the *lldD2* gene, we devised a metric for transmissibility associated with each mutation akin to (*6*). First, homoplasies in this gene were categorized as being in the promotor region, in codons 3 and 253. All homoplasy emergences were mapped to the respective branch in the full phylogenetic time tree. For each homoplasy, we recorded the subtree's number of descendents and number of deme transitions. The reasoning behind this was that ancestors with homoplasic mutations increasing transmissibility should have more descendents and more deme transitions per time than ancestors without these mutations. For this latter control group, we randomly extracted subtrees of comparable height distribution to those subtrees with promoter/codon 3/codon 253 mutations, and we labeled this control group as "none" (that is, no homoplasic mutations in *lldD2*). Figure S2 shows a linear model between subtree height and the number of sampled descendents by each mutation category. To test whether these categories have different slopes, we used an analysis of covariance (ANCOVA) procedure. First, a simple null model, where the number of descendents are dependent on subtree height but not on the mutation category (including the none group), was set up. An alternative model is that the relationship between the subtree height and the number of descendents varies between these four different mutation categories. For each of these models, we weighted the number of deme transitions as (number of deme transmissions)^2 + 1, i.e., no transitions got a weight of 1, one transition got a weight of 2, and two transitions got a weight of 5. The ANCOVA rejected the null model, showing significant preference for the per-group model (F test, $P = 0.038$). Analysis of individual height-group interaction terms showed that the coefficient for promotor mutations was significantly different from zero, indicating a positive association between *lldD2* promotor mutations and transmissibility. Note that, if the weighting by deme transitions is removed, the ANCOVA no longer significantly prefers the alternative model (F test, $P = 0.114$), and there is no evidence for homoplasy group interaction with height and number of descendants.

Phylogenetic and phylogeographic inference

To estimate substitution rates by means of sampling date calibrated Bayesian evolutionary analyses (*52*), we down-sampled the collection to a manageable size by including a maximum of 20 randomly chosen isolates from each country. This resulted in a sample collection of 269 genomes. Three ancient *M.tb* genomes from 18th century Hungarian mummies (*37*) and five Danish isolates from the 1960s (*38*) were included to provide temporal structure to the data.

A total of 7994 variable sites remained after selecting the 269 genomes. A preliminary check using TempEst (*53*) confirmed a moderate but highly significant temporal signal in the data (fig. S4), which was also confirmed by tip randomization (see below). A GTR substitution model was chosen on the basis of model testing, as described above. On the basis of marginal-likelihood estimation (MLE), an exponential demographic model was found to best fit the data [tested against the constant population size and Gaussian Markov Random Fields (GMRF) skyride models (*54*)]. Also based on MLE, an uncorrelated relaxed clock was favored over a strict clock model. Three independent BEAST Markov Chain Monte Carlo (MCMC) chains were run, and convergence to a stationary posterior distribution was confirmed both within and between chains. These analyses resulted in an estimated substitution rate of 4.84×10^{-8} (95% HPD, 4.16×10^{-8} to 5.44×10^{-8}) substitutions per site per year and an estimated time of the MRCA in 1096 CE (95% HPD, 955 to 1231). To assess the robustness of the temporal inference, we performed 10 additional runs after randomization of the sampling dates (*55*). None of the randomized runs had rate estimates overlapping with the inference using real sampling dates (fig. S3), supporting the robustness of the original inference.

We then ran a larger dataset (1207 genomes, after retaining a maximum of five representatives from each of three densely sampled outbreaks from Argentina and 10 from Quebec, Canada) in BEAST v1.8.4 (*56*) with a fixed substitution rate as estimated above. The MRCA of this tree was inferred to have existed in the year 1157 (HPD intervals not reported due to our application of a fixed rate in this analysis). The overall accuracy of the dating inferences in BEAST was further supported by independent analyses applying Least Squares Dating (LSD) v0.3 (*57*), resulting in an estimated time to MRCA in 1195 CE (95% HPD, 1061 to 1270) for the 1207 genome dataset. The tree generated for 1207 genomes in BEAST v1.8.4 was used as an input tree for phylogeographic analyses using both the BASTA module in BEAST2 and simple DTA as implemented in BEAST v1.8.4 (*9*).

To reduce computational complexity and increase post-analysis interpretability, we collapsed the country of origin information into five distinct UN regions: North America (United States and Canada), South America (Brazil, Argentina, and Peru), Africa (Democratic Republic of the Congo, Malawi, South Africa, and Uganda), Europe (Denmark, the Netherlands, Hungary, Portugal, Russia, and United Kingdom), and Southeast Asia (Vietnam).

In the discrete trait model (*7*), we used a GTR model, restricted the clock rate to 4.84×10^{-8}, and set the starting tree as specified above. The prior population size was set as constant. A symmetric deme substitution model was used, and we turned Bayesian stochastic search variable selection on. Ancestral states were reconstructed at all nodes. Other priors were left at default values. The chain was run for 10,000,000 generations with logging every 10,000th iteration, and three of these runs were combined to create the final posterior sample of trees and parameters. We verified chain convergence and good mixing and an effective sample size (ESS) > 200 for all parameters using Tracer (http://tree.bio.ed.ac.uk/software/tracer/). A maximum clade

credibility (MCC) tree was created using TreeAnnotator (http://beast.community/treeannotator), with 20% of the chain discarded as burn in. The resulting tree displayed a European root with 92% posterior probability.

A separate phylogeographic reconstruction of the L4.3 RdRio family was also performed. This was completed by manually extracting the RdRio subtree from the full time tree and then setting up a new DTA run consisting of these 243 isolates alone. For this analysis, we acquired patient country of origin data for genomes from our Portuguese and Dutch collections, which led to a changed country of origin for 13 genomes in this dataset. To restrict the number of possible demes of low/single sample size, we collapsed countries into eight different geographic categories: Iberia (Portugal, Spain, and Cape Verde), North Europe (United Kingdom, Russia, Bosnia, Netherlands, and Germany), Peru, Atlantic South America (Venezuela, Brazil, Aruba, and Suriname), Congo/Angola, Malawi/Mozambique, Uganda, and South Africa. The prior population size was set to 10,000. Other than this, the analysis was run with the same parameters as for the full dataset DTA. The rationale for placing Cape Verde in the Iberia category is that Cape Verde was uninhabited before Portuguese settlement in the 15th century.

As a complement to DTA, we used the BEAST2 module BASTA v2.3.1 (*8*) for phylogeographic inferences. We specified a migration model with the same five demes as above. The initial values for deme transition rate were set to 1.0×10^{-3} and the subpopulation to 6000, and these numbers correspond to the median outputs from DTA. The rate matrix and population size priors were given log-normal prior distributions with $M = -10$ and $S = 2.0$ and $M = 9.0$ and $S = 0.6$, respectively. Since it was not possible to place deme restrictions on internal nodes in BASTA, we artificially introduced an isolate with a branch length of 1.0×10^{-10} from the root, and the location of this isolate was set to correspond to each of the five demes in different runs. The results of these five runs were subsequently evaluated jointly. We ran each chain for 1,000,000 generations with storing set to every 1000th iteration. We disabled all scaling operators except the rate scaler, which was given a scale factor of 0.8 and a weight of 1.0, and the population size scaler, which was given a weight of 3 and a scale factor of 0.8, and degrees of freedom was set to 1 (subpopulation sizes effectively set to equal). Since we knew that the MRCA of all isolates existed roughly around the year 1100 CE, before European colonization of the Americas, we discarded all trees with a root inferred to be from North or South America. The parameter logs were inspected in the same way as described for DTA, and an MCC tree was made using TreeAnnotator v.2.4.5, with burn in set to 20% and node heights set to median. All BEAST and BEAST2 runs were run locally or on the Cyberinfrastructure for phylogenetic research (CIPRES) science gateway (*58*).

Migration matrices (*10*) were constructed to visualize the overall patterns of migration inferred by the two methods. Both methods inferred Europe to have played a pivotal role in sourcing the rest of the world with TB (fig. S5).

To study migration over time, conceptually mirroring the methodology used in (*10*), we wrote an in-house script (available at https://github.com/admiralenola/globall4scripts) to read the MCC tree from the BASTA runs using the ETE toolkit (*59*) and traversed the time tree in a sliding window fashion, for each year writing out the number of branches corresponding to the five different demes. In our analyses, migration events were set to occur on nodes but, in reality, could have occurred at any point along the branch downstream of this node. This introduces a slight bias toward inflated ages of migration events, which is most pronounced for very early migration events but negligible for later migrations due to extensive branching.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/10/eaat5869/DC1

Supplementary Text

Fig. S1. Genome-wide assessment of homoplasic mutations presented as gene-wise parsimony scores as a function of θ and nucleotide diversity (π).

Fig. S2. Node descendants as a function of node age for clones harboring different groups of *lldD2* mutations.

Fig. S3. Tip-randomization results.

Fig. S4. Root-to-tip analysis performed on a global down-sampled collection containing 269 genomes.

Fig. S5. Migration matrices inferred with DTA and BASTA visualized as heatmaps.

Fig. S6. Inferred migration of L4 over time from Europe to North and South America, as well as within the continents, using BASTA.

Fig. S7. Phylogeographic reconstruction of the RdRio family.

Fig. S8. Full temporal phylogeny of L4 including node age 95% HPD intervals.

Fig. S9. Histogram summarizing the emergence of *lldD2* mutations over time.

Dataset S1. List of strains.

Dataset S2. Results from population genomic inferences.

Dataset S3. Location of lldD2 mutation emergence.

References (*65*–*66*)

REFERENCES AND NOTES

1. I. Comas, M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites, D. Yeboah-Manu, G. Bothamley, J. Mei, L. Wei, S. Bentley, S. R. Harris, S. Niemann, R. Diel, A. Aseffa, Q. Gao, D. Young, S. Gagneux, Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. Genet.* **45**, 1176–1182 (2013).
2. K. I. Bos, K. M. Harkins, A. Herbig, M. Coscolla, N. Weber, I. Comas, S. A. Forrest, J. M. Bryant, S. R. Harris, V. J. Schuenemann, T. J. Campbell, K. Majander, A. K. Wilbur, R. A. Guichon, Dawnie L. Wolfe Steadman, D. C. Cook, S. Niemann, M. A. Behr, M. Zumarraga, R. Bastida, D. Huson, K. Nieselt, D. Young, J. Parkhill, J. E. Buikstra, S. Gagneux, A. C. Stone, J. Krause, Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–497 (2014).
3. F. Coll, R. McNerney, J. A. Guerra-Assunção, J. R. Glynn, J. Perdigão, M. Viveiros, I. Portugal, A. Pain, N. Martin, T. G. Clark, A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* **5**, 4812 (2014).
4. D. Stucki, D. Brites, L. Jeljeli, M. Coscolla, Q. Liu, A. Trauner, L. Fenner, L. Rutaihwa, S. Borrell, T. Luo, Q. Gao, M. Kato-Maeda, M. Ballif, M. Egger, R. Macedo, H. Mardassi, M. Moreno, G. T. Vilanova, J. Fyfe, M. Globan, J. Thomas, F. Jamieson, J. L. Guthrie, A. Asante-Poku, D. Yeboah-Manu, E. Wampande, W. Ssengooba, M. Joloba, W. H. Boom, I. Basu, J. Bower, M. Saraiva, S. E. G. Vasconcellos, P. Suffys, A. Koch, R. Wilkinson, L. Gail-Bekker, B. Malla, S. D. Ley, H.-P. Beck, B. C. de Jong, K. Toit, E. Sanchez-Padilla, M. Bonnet, A. Gil-Brusola, M. Frank, V. N. Penlap Beng, K. Eisenach, I. Alani, Pe. W. Ndung'u, G. Revathi, F. Gehre, S. Akter, F. Ntoumi, L. Stewart-Isherwood, N. E. Ntinginya, A. Rachow, M. Hoelscher, D. M. Cirillo, G. Skenders, S. Hoffner, D. Bakonyte, P. Stakenas, R. Diel, V. Crudu, O. Moldovan, S. Al-Hajoj, L. Otero, F. Barletta, E. J. Carter, L. Diero, P. Supply, I. Comas, S. Niemann, S. Gagneux, *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48**, 1535–1543 (2016).
5. J. A. Guerra-Assunção, A. C. Crampin, R. M. Houben, T. Mzembe, K. Mallard, F. Coll, P. Khan, L. Banda, A. Chiwaya, R. P. Pereira, R. McNerney, P. E. Fine, J. Parkhill, T. G. Clark, J. R. Glynn, Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife* **4** 10.7554/eLife.05166 (2015).
6. K. E. Holt, P. McAdam, V. K. T. Phan, T. M. H. Dang, N. L. Nguyen, H. L. Nguyen, T. Q. N. Nguyen, T. T. T. Nguyen, G. Thwaites, D. J. Edwards, K. Pham, J. Farrar, C. C. Khor, Y. Y. Teo, M. Inouye, M. Caws, S. J. Dunstan, Genomic analysis of Mycobacterium tuberculosis reveals complex etiology of tuberculosis in Vietnam including frequent introduction and transmission of Beijing lineage and positive selection for EsxW Beijing variant. *bioRxiv* 10.1101/092189 (2016).
7. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots. *PLOS Comput. Biol.* **5**, e1000520 (2009).
8. N. De Maio, C.-H. Wu, K. M. O'Reilly, D. Wilson, New routes to phylogeography: A Bayesian structured coalescent approximation. *PLOS Genet.* **11**, e1005421 (2015).
9. A. J. Drummond, M. A. Suchard, D. Xie, A. Rambaut, Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).

10. M. B. O'Neill, A. C. Shockey, A. Zarley, W. Aylward, V. Eldholm, A. Kitchen, C. S. Pepperell, Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia. bioRxiv 10.1101/210161. 27 October 2017.
11. O. Chapuis, *A History of Vietnam: From Hong Bang to Tu Duc* (Greenwood Press, 1995).
12. H. D. Donoghue, Paleomicrobiology of human tuberculosis. *Microbiol. Spectr.* **4**, 10.1128/microbiolspec.PoH-0003-2014 (2016).
13. V. Ritacco, B. López, P. I. Cafrune, L. Ferrazoli, P. N. Suffys, N. Candia, L. Vásquez, T. Realpe, J. Fernández, K. V. Lima, J. Zurita, J. Robledo, M. L. Rossetti, A. L. Kritski, M. A. Telles, J. C. Palomino, H. Heersma, D. van Soolingen, K. Kremer, L. Barrera, Mycobacterium tuberculosis strains of the Beijing genotype are rarely observed in tuberculosis patients in South America. *Mem. Inst. Oswaldo Cruz* **103**, 489–492 (2008).
14. N. D. Cook, *Born to Die - Disease and New World Conquest, 1492-1650* (Cambridge Univ. Press, 1998), vol. 1.
15. K. A. Cohen, T. Abeel, A. Manson McGuire, C. A. Desjardins, V. Munsamy, T. P. Shea, B. J. Walker, N. Bantubani, D. V. Almeida, L. Alvarado, S. B. Chapman, N. R. Mvelase, E. Y. Duffy, M. G. Fitzgerald, P. Govender, S. Gujja, S. Hamilton, C. Howarth, J. D. Larimer, K. Maharaj, M. D. Pearson, M. E. Priest, Q. Zeng, N. Padayatchi, J. Grosset, S. K. Young, J. Wortman, K. P. Mlisana, M. R. O'Donnell, B. W. Birren, W. R. Bishai, A. S. Pym, A. M. Earl, Evolution of extensively drug-resistant tuberculosis over four decades: Whole genome sequencing and dating analysis of mycobacterium tuberculosis isolates from KwaZulu-Natal. *PLOS Med.* **12**, e1001880 (2015).
16. N. S. Shah, S. C. Auld, J. C. Brust, B. Mathema, N. Ismail, P. Moodley, K. Mlisana, S. Allana, A. Campbell, T. Mthiyane, N. Morris, P. Mpangase, H. van der Meulen, S. V. Omar, T. S. Brown, A. Narechania, E. Shaskina, T. Kapwata, B. Kreiswirth, N. R. Gandhi, Transmission of extensively drug-resistant tuberculosis in South Africa. *N. Engl. J. Med.* **376**, 243–253 (2017).
17. F. Lanzas, P. C. Karakousis, J. C. Sacchettini, T. R. Ioerger, Multidrug-resistant tuberculosis in panama is driven by clonal expansion of a multidrug-resistant Mycobacterium tuberculosis strain related to the KZN extensively drug-resistant M. tuberculosis strain from South Africa. *J. Clin. Microbiol.* **51**, 3277–3285 (2013).
18. L. C. O. Lazzarini, R. C. Huard, N. L. Boechat, H. M. Gomes, M. C. Oelemann, N. Kurepina, E. Shashkina, F. C. Mello, A. L. Gibson, M. J. Virginio, A. G. Marsico, W. R. Butler, B. N. Kreiswirth, P. N. Suffys, J. R. Lappa E Silva, J. L. Ho, Discovery of a novel *Mycobacterium tuberculosis* lineage that is a major cause of tuberculosis in Rio de Janeiro, Brazil. *J. Clin. Microbiol.* **45**, 3891–3902 (2007).
19. S. A. Weisenberg, A. L. Gibson, R. C. Huard, N. Kurepina, H. Bang, L. C. Lazzarini, Y. Chiu, J. Li, S. Ahuja, J. Driscoll, B. N. Kreiswirth, J. L. Ho, Distinct clinical and epidemiological features of tuberculosis in New York City caused by the RD(Rio) Mycobacterium tuberculosis sublineage. *Infect. Genet. Evol.* **12**, 664–670 (2012).
20. S. David, E. L. Duarte, C. Q. Leite, J. N. Ribeiro, J. N. Maio, E. Paixão, C. Portugal, L. Sancho, J. Germano de Sousa, Implication of the RD(Rio) Mycobacterium tuberculosis sublineage in multidrug resistant tuberculosis in Portugal. *Infect. Genet. Evol.* **12**, 1362–1367 (2012).
21. V. Eldholm, J. Monteserin, A. Rieux, B. Lopez, B. Sobkowiak, V. Ritacco, F. Balloux, Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat. Commun.* **6**, 7119 (2015).
22. V. Ritacco, B. López, M. Ambroggi, D. Palmero, B. Salvadores, E. Gravina; National TB Laboratory Network, S. Imaz, L. Barrera, HIV infection and geographically bound transmission of drug-resistant tuberculosis, Argentina. *Emerging Infect. Dis.* **18**, 1802–1810 (2012).
23. C. S. Pepperell, J. M. Granka, D. C. Alexander, M. A. Behr, L. Chui, J. Gordon, J. L. Guthrie, F. B. Jamieson, D. Langlois-Klassen, R. Long, D. Nguyen, W. Wobeser, M. W. Feldman, Dispersal of Mycobacterium tuberculosis via the Canadian fur trade. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 6526–6531 (2011).
24. M. R. Farhat, B. J. Shapiro, K. J. Kieser, R. Sultana, K. R. Jacobson, T. C. Victor, R. M. Warren, E. M. Streicher, A. Calver, A. Sloutsky, D. Kaur, J. E. Posey, B. Plikaytis, M. R. Oggioni, J. L. Gardy, J. C. Johnston, M. Rodrigues, P. K. Tang, M. Kato-Maeda, M. L. Borowsky, B. Muddukrishna, B. N. Kreiswirth, N. Kurepina, J. Galagan, S. Gagneux, B. Birren, E. J. Rubin, E. S. Lander, P. C. Sabeti, M. Murray, Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **45**, 1183–1189 (2013).
25. S. Billig, M. Schneefeld, C. Huber, G. A. Grassl, W. Eisenreich, F. C. Bange, Lactate oxidation facilitates growth of Mycobacterium tuberculosis in human macrophages. *Sci. Rep.* **7**, 6484 (2017).
26. N. S. Osório, F. Rodrigues, S. Gagneux, J. Pedrosa, M. Pinto-Carbó, A. G. Castro, D. Young, I. Comas, M. Saraiva, Evidence for diversifying selection in a set of Mycobacterium tuberculosis genes in response to antibiotic- and nonantibiotic-related pressure. *Mol. Biol. Evol.* **30**, 1326–1336 (2013).
27. K. Dheda, T. Gumbo, G. Maartens, K. E. Dooley, R. McNerney, M. Murray, J. Furin, E. A. Nardell, L. London, E. Lessem, G. Theron, P. van Helden, S. Niemann, M. Merker, D. Dowdy, A. Van Rie, G. K. Siu, J. G. Pasipanodya, C. Rodrigues, T. G. Clark, F. A. Sirgel, A. Esmail, H. H. Lin, S. R. Atre, H. S. Schaaf, K. C. Chang, C. Lange, P. Nahid, Z. F. Udwadia, C. R. Horsburgh Jr., G. J. Churchyard, D. Menzies, A. C. Hesseling, E. Nuermberger, H. McIlleron, K. P. Fennelly, E. Goemaere, E. Jaramillo, M. Low, C. M. Jara, N. Padayatchi, R. M. Warren, The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *Lancet Respir. Med.* (2017).
28. A. L. Manson, K. A. Cohen, T. Abeel, C. A. Desjardins, D. T. Armstrong, C. E. Barry III; TBResist Global Genome Consortium, S. B. Chapman, S. N. Cho, A. Gabrielian, J. Gomez, A. M. Jodals, M. Joloba, P. Jureen, J. S. Lee, L. Malinga, M. Maiga, D. Nordenberg, E. Noroc, E. Romancenco, A. Salazar, W. Ssengooba, A. A. Velayati, K. Winglee, A. Zalutskaya, L. E. Via, G. H. Cassell, S. E. Dorman, J. Ellner, P. Farnia, J. E. Galagan, A. Rosenthal, V. Crudu, D. Homorodean, P. R. Hsueh, S. Narayanan, A. S. Pym, A. Skrahina, S. Swaminathan, M. Van der Walt, D. Alland, W. R. Bishai, T. Cohen, S. Hoffner, B. W. Birren, A. M. Earl, Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nat. Genet.* **49**, 395–402 (2017).
29. V. Eldholm, F. Balloux, Antimicrobial resistance in mycobacterium tuberculosis: The odd one out. *Trends Microbiol.* **24**, 637–648 (2016).
30. N. Casali, N. Casali, V. Nikolayevskyy, Y. Balabanova, S. R. Harris, O. Ignatyeva, I. Kontsevaya, J. Corander, J. Bryant, J. Parkhill, S. Nejentsev, R. D. Horstmann, T. Brown, F. Drobniewski, Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat. Genet.* **46**, 279–286 (2014).
31. R. S. Lee, N. Radomski, J. F. Proulx, I. Levade, B. J. Shapiro, F. McIntosh, H. Soualhine, D. Menzies, M. A. Behr, Population genomics of Mycobacterium tuberculosis in the Inuit. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13609 (2015).
32. S. Malm, L. S. G. Linguissi, E. M. Tekwu, J. C. Vouvoungui, T. A. Kohl, P. Beckert, A. Sidibe, S. Rüsch-Gerdes, I. K. Madzou-Laboum, S. Kwedi, V. P. Beng, M. Frank, F. Ntoumi, S. Niemann, New *Mycobacterium tuberculosis* complex sublineage, Brazzaville, Congo. *Emerging Infect. Dis.* **23**, 423 (2017).
33. J. R. Glynn, J. A. Guerra-Assunção, R. M. Houben, L. Sichali, T. Mzembe, L. K. Mwaungulu, J. N. Mwaungulu, R. McNerney, P. Khan, J. Parkhill, A. C. Crampin, T. G. Clark, Whole genome sequencing shows a low proportion of tuberculosis disease is attributable to known close contacts in rural Malawi. *PLOS ONE* **10**, e0132840 (2015).
34. J. M. Bryant, A. C. Schürch, H. van Deutekom, S. R. Harris, J. L. de Beer, V. de Jager, K. Kremer, S. A. van Hijum, R. J. Siezen, M. Borgdorff, S. D. Bentley, J. Parkhill, D. van Soolingen, Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infect. Dis.* **13**, 110 (2013).
35. J. Perdigao, H. Silva, D. Machado, R. Macedo, F. Maltez, C. Silva, L. Jordao, I. Couto, K. Mallard, F. Coll, G. A. Hill-Cawthorne, R. McNerney, A. Pain, T. G. Clark, M. Viveiros, I. Portugal, Unraveling *Mycobacterium tuberculosis* genomic diversity and evolution in Lisbon, Portugal, a highly drug resistant setting. *BMC Genomics* **15**, 991 (2014).
36. T. S. Brown, A. Narechania, J. R. Walker, P. J. Planet, P. J. Bifani, S. O. Kolokotronis, B. N. Kreiswirth, B. Mathema, Genomic epidemiology of Lineage 4 Mycobacterium tuberculosis subpopulations in New York City and New Jersey, 1999-2009. *BMC Genomics* **17**, 947 (2016).
37. G. L. Kay, M. J. Sergeant, Z. Zhou, J. Z. Chan, A. Millard, J. Quick, I. Szikossy, I. Pap, M. Spigelman, N. J. Loman, M. Achtman, H. D. Donoghue, M. J. Pallen, Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat. Commun.* **6**, 6717 (2015).
38. T. Lillebaek, A. Norman, E. M. Rasmussen, R. L. Marvig, D. B. Folkvardsen, Å. B. Andersen, L. Jelsbak, Substantial molecular evolution and mutation rates in prolonged latent Mycobacterium tuberculosis infection in humans. *Int. J. Med. Microbiol.* **306**, 580–585 (2016).
39. V. Eldholm, J. H.-O. Pettersson, O. B. Brynildsrud, A. Kitchen, E. M. Rasmussen, T. Lillebaek, J. O. Rønning, V. Crudu, A. T. Mengshoel, N. Debech, K. Alfsnes, J. Bohlin, C. S. Pepperell, F. Balloux, Armed conflict and population displacement as drivers of the evolution and dispersal of *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 13881–13886 (2016).
40. F. Coll, R. McNerney, M. D. Preston, J. A. Guerra-Assunção, A. Warry, G. Hill-Cawthorne, K. Mallard, M. Nair, A. Miranda, A. Alves, J. Perdigão, M. Viveiros, I. Portugal, Z. Hasan, R. Hasan, J. R. Glynn, N. Martin, A. Pain, T. G. Clark, Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **7**, 51 (2015).
41. S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, B. G. Barrell, Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537–544 (1998).
42. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

43. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
44. B. Q. Minh, M. A. T. Nguyen, A. von Haeseler, Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
45. S. B. Georghiou, M. Magana, R. S. Garfein, D. G. Catanzaro, A. Catanzaro, T. C. Rodwell, Evaluation of genetic mutations associated with Mycobacterium tuberculosis resistance to amikacin, kanamycin and capreomycin: A systematic review. *PLOS ONE* **7**, e33275 (2012).
46. M. R. Farhat, K. R. Jacobson, M. F. Franke, D. Kaur, A. Sloutsky, C. D. Mitnick, M. Murray, Gyrase mutations are associated with variable levels of fluoroquinolone resistance in *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* **54**, 727–733 (2016).
47. S. Malik, M. Willby, D. Sikes, O. V. Tsodikov, J. E. Posey, New insights into fluoroquinolone resistance in *Mycobacterium tuberculosis*: Functional genetic analysis of *gyrA* and *gyrB* mutations. *PLOS ONE* **7**, e39754 (2012).
48. P. L. Rice, A. Bleasby, I. Longden, EMBOSS: The European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
49. I. Letunic, P. Bork, Interactive Tree Of Life v2: Online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).
50. S. De Mita, M. Siol, EggLib: Processing, analysis and simulation tools for population genetics and genomics. *BMC Genet.* **13**, 27 (2012).
51. J. Sukumaran, M. T. Holder, DendroPy: A Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).
52. A. J. Drummond, G. Nicholls, A. G. Rodrigo, W. Solomon, Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307–1320 (2002).
53. A. Rambaut, T. T. Lam, L. M. Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
54. V. N. Minin, E.W. Bloomquist, M. A. Suchard, Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459–1471 (2008).
55. A. Rieux, C. E. Khatchikian, TIPDATINGBEAST: An R package to assist the implementation of phylogenetic tip-dating tests using beast. *Mol. Ecol. Resour.* **17**, 608–613 (2017).
56. A. J. Drummond, A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
57. T.-H. To, M. Jung, S. Lycett, O. Gascuel, Fast dating using least-squares criteria and algorithms. *Syst. Biol.* **65**, 82–97 (2016).
58. M. Miller, W. Pfeiffer, T. Schwartz, Creating the CIPRES Science Gateway for inference of large phylogenetic trees, in *Proceedings of the Gateway Computing Environments Workshop (GCE)*, New Orleans, Louisiana, USA, 14 November 2010, pp. 1–8.
59. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
60. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
61. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 (2013).
62. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 (2012).
63. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin; 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
64. P. Cingolani, A. Platts, L. le Wang, M. Coon, T Nguyen, L Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
65. D. Nguyen, P. Brassard, J. Westley, L. Thibert, M. Proulx, K. Henry, K. Schwartzman, D. Menzies, M. A. Behr, Widespread pyrazinamide-resistant Mycobacterium tuberculosis family in a low-incidence setting. *J. Clin. Microbiol.* **41**, 2878–2883 (2003).
66. M. A. Behr, W. R. Waters, Is tuberculosis a lymphatic disease with a pulmonary portal? *Lancet Infect. Dis.* **14**, 250–255 (2014).

Acknowledgments: We are grateful to N. de Maio for helpful tips with BASTA. We would also like to acknowledge the Norwegian Sequencing Centre for their efficient and competent handling of our sequencing needs. The Snippy pipeline used for SNP calling relied on the following independent code and software: SAMtools (*60*), BWA (*61*), FreeBayes (*62*), Vcflib (github.com/ekg/vcflib), VCFtools (*63*), and SnpEff (*64*). **Funding:** C.S.P. was supported by the NIH grant number 1R01AI113287-01A1. K.E.H. is funded by a Viertel Foundation of Australia Senior Medical Research Fellowship. F.B. acknowledges support from the BBSRC GCRF scheme and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. **Author contributions:** V.E. and O.B.B. conceived the study. Ideas and methodology were developed by V.E., O.B.B., C.S.P., F.B., E.J.F., B.M., K.E.H., and M.B.O. Data were generated or made available by P.S., C.S.P., L.G., J.M., V.R., T.C., N.D., I.K., T.S.B., D.v.S., F.F., M.A.d.S., J.P., I.P., M.V., S.D., P.V.K.T., B.L., B.M., M.C., and K.A. Analyses were performed by O.B.B., V.E., M.B.O., J.B., K.A., and J.O.-H.P. The paper was drafted by V.E. and O.B.B. and finalized with the aid of C.S.P., F.B., E.J.F., K.E.H., M.B.O., B.M., P.S., L.G., and A.K. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 14 March 2018
Accepted 11 September 2018
Published 17 October 2018
10.1126/sciadv.aat5869

Citation: O. B. Brynildsrud, C. S. Pepperell, P. Suffys, L. Grandjean, J. Monteserin, N. Debech, J. Bohlin, K. Alfsnes, J. O.-H. Pettersson, I. Kirkeleite, F. Fandinho, M. A. da Silva, J. Perdigao, I. Portugal, M. Viveiros, T. Clark, M. Caws, S. Dunstan, P. V. K. Thai, B. Lopez, V. Ritacco, A. Kitchen, T. S. Brown, D. van Soolingen, M. B. O'Neill, K. E. Holt, E. J. Feil, B. Mathema, F. Balloux, V. Eldholm, Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Sci. Adv.* **4**, eaat5869 (2018).

Science Advances

Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation

Ola B. Brynildsrud, Caitlin S. Pepperell, Philip Suffys, Louis Grandjean, Johana Monteserin, Nadia Debech, Jon Bohlin, Kristian Alfsnes, John O.-H. Pettersson, Ingerid Kirkeleite, Fatima Fandinho, Marcia Aparecida da Silva, Joao Perdigao, Isabel Portugal, Miguel Viveiros, Taane Clark, Maxine Caws, Sarah Dunstan, Phan Vuong Khac Thai, Beatriz Lopez, Viviana Ritacco, Andrew Kitchen, Tyler S. Brown, Dick van Soolingen, Mary B. O'Neill, Kathryn E. Holt, Edward J. Feil, Barun Mathema, Francois Balloux and Vegard Eldholm

Sci Adv **4** (10), eaat5869.
DOI: 10.1126/sciadv.aat5869