Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa

Comas, Iñaki; Hailu, Elena; Kiros, Teklu; Bekele, Shiferaw; Mekonnen, Wondale; Gumi, Balako; Tschopp, Rea; Ameni, Gobena; Hewinson, R. Glyn; Robertson, Brian D.; Goig, Galo A.; Stucki, David; Gagneux, Sebastien; Aseffa, Abraham; Young, Douglas; Berg, Stefan

Published in:
Current Biology

DOI:
10.1016/j.cub.2015.10.061

Publication date:
2015

Citation for published version (APA):
Comas, I., Hailu, E., Kiros, T., Bekele, S., Mekonnen, W., Gumi, B., Tschopp, R., Ameni, G., Hewinson, R. G., Robertson, B. D., Goig, G. A., Stucki, D., Gagneux, S., Aseffa, A., Young, D., & Berg, S. (2015). Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa. *Current Biology*, *25*(24), 3260-3266. https://doi.org/10.1016/j.cub.2015.10.061

Report

Current Biology

Population Genomics of *Mycobacterium tuberculosis* in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa

Highlights

- Ethiopia hosts a complex population structure of *Mycobacterium tuberculosis*
- Our analyses provide further support for an African origin for tuberculosis
- Our analyses discard the "virgin soil" hypothesis of tuberculosis in Sub-Saharan Africa

Authors

Iñaki Comas, Elena Hailu, Teklu Kiros, ..., Abraham Aseffa, Douglas Young, Stefan Berg

Correspondence

inaki.comas@uv.es (I.C.), stefan.berg@apha.gsi.gov.uk (S.B.)

In Brief

Comas et al. highlight a complex population structure of *Mycobacterium tuberculosis* in Ethiopia through genome sequencing. Analyses reveal a mixture of sub-lineages with global distribution as well as those specific to Africa, providing further support for an African origin for tuberculosis and contradicting the earlier "virgin soil" hypothesis.

Comas et al., 2015, Current Biology *25*, 3260–3266
December 21, 2015
http://dx.doi.org/10.1016/j.cub.2015.10.061

CellPress

CellPress

Current Biology
Report

Population Genomics of *Mycobacterium tuberculosis* in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa

Iñaki Comas,[1,2,*] Elena Hailu,[3] Teklu Kiros,[3] Shiferaw Bekele,[3] Wondale Mekonnen,[3] Balako Gumi,[3] Rea Tschopp,[3,4] Gobena Ameni,[5] R. Glyn Hewinson,[6] Brian D. Robertson,[7] Galo A. Goig,[1] David Stucki,[8] Sebastien Gagneux,[8] Abraham Aseffa,[3] Douglas Young,[9] and Stefan Berg[6,*]

[1]Genomics and Health Unit, FISABIO Public Health, Valencia 46020, Spain
[2]CIBER (Centros de Investigación Biomédica en Red) in Epidemiology and Public Health, Instituto de Salud Carlos III, Madrid 28029, Spain
[3]Armauer Hansen Research Institute, PO Box 1005, Addis Ababa, Ethiopia
[4]Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Basel 4002, and University of Basel, Basel 4003, Switzerland
[5]Aklilu Lemma Institute of Pathobiology, Addis Ababa University, PO Box 1176, Addis Ababa, Ethiopia
[6]Bovine TB Research Group, Animal and Plant Health Agency, Surrey KT15 3NB, UK
[7]Center for Molecular Bacteriology and Infection, Department of Medicine, Flowers Building, South Kensington, Imperial College London, London SW7 2AZ, UK
[8]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel 4002, and University of Basel, Basel 4003, Switzerland
[9]The Francis Crick Institute, Mill Hill Laboratory, The Ridgeway, Mill Hill, London NW7 1AA, UK
*Correspondence: inaki.comas@uv.es (I.C.), stefan.berg@apha.gsi.gov.uk (S.B.)
http://dx.doi.org/10.1016/j.cub.2015.10.061

SUMMARY

Colonial medical reports claimed that tuberculosis (TB) was largely unknown in Africa prior to European contact, providing a "virgin soil" for spread of TB in highly susceptible populations previously unexposed to the disease [1, 2]. This is in direct contrast to recent phylogenetic models which support an African origin for TB [3–6]. To address this apparent contradiction, we performed a broad genomic sampling of *Mycobacterium tuberculosis* in Ethiopia. All members of the *M. tuberculosis* complex (MTBC) arose from clonal expansion of a single common ancestor [7] with a proposed origin in East Africa [3, 4, 8]. Consistent with this proposal, MTBC lineage 7 is almost exclusively found in that region [9–11]. Although a detailed medical history of Ethiopia supports the view that TB was rare until the 20th century [12], over the last century Ethiopia has become a high-burden TB country [13]. Our results provide further support for an African origin for TB, with some genotypes already present on the continent well before European contact. Phylogenetic analyses reveal a pattern of serial introductions of multiple genotypes into Ethiopia in association with human migration and trade. In place of a "virgin soil" fostering the spread of TB in a previously naive population, we propose that increased TB mortality in Africa was driven by the introduction of European strains of *M. tuberculosis* alongside expansion of selected indigenous strains having biological characteristics that carry a fitness benefit in the urbanized settings of post-colonial Africa.

RESULTS AND DISCUSSION

Population Structure of *M. tuberculosis* in Ethiopia

The high incidence of tuberculosis (TB) among Africans during European colonization in the late 19th and early 20th centuries gave rise to the hypothesis that Africa was a "virgin soil," particularly permissive to the spread of the disease in previously unexposed populations [1, 2]. This stands in sharp contrast to more recent evolutionary models that propose an African origin for *Mycobacterium tuberculosis* [3–5]. To address the apparent contradiction between an African origin of human TB and the "virgin soil" hypothesis, we performed a broad genomic sampling of *M. tuberculosis* in Ethiopia. We sequenced 66 strains from a previously genotyped *M. tuberculosis* collection [9], selected to represent the most common spoligotypes found in Ethiopia along with examples of rare genotypic outliers and high-density sampling of the unique Ethiopian lineage 7 (L7) (Table S1). Figure 1 illustrates the phylogenetic structure of these Ethiopian isolates in the context of a previously described panel of 219 *M. tuberculosis* complex (MTBC) strains representative of the global diversity [3]. In addition to L7, the Ethiopian isolates belonged to three *M. tuberculosis* lineages: lineage 1 (L1), commonly associated with populations living around the Indian Ocean; lineage 3 (L3), common in Central Asia but also prevalent in East Africa; and lineage 4 (L4), the widespread Euro-American lineage [14]. Mapping to a recent SNP-based classification system [15] identified representatives of five of the eight L4 sub-lineages in the Ethiopian genome dataset. Within L3, we were able to assign Ethiopian strains to three of the five defined sub-lineages along with novel sub-lineages that we

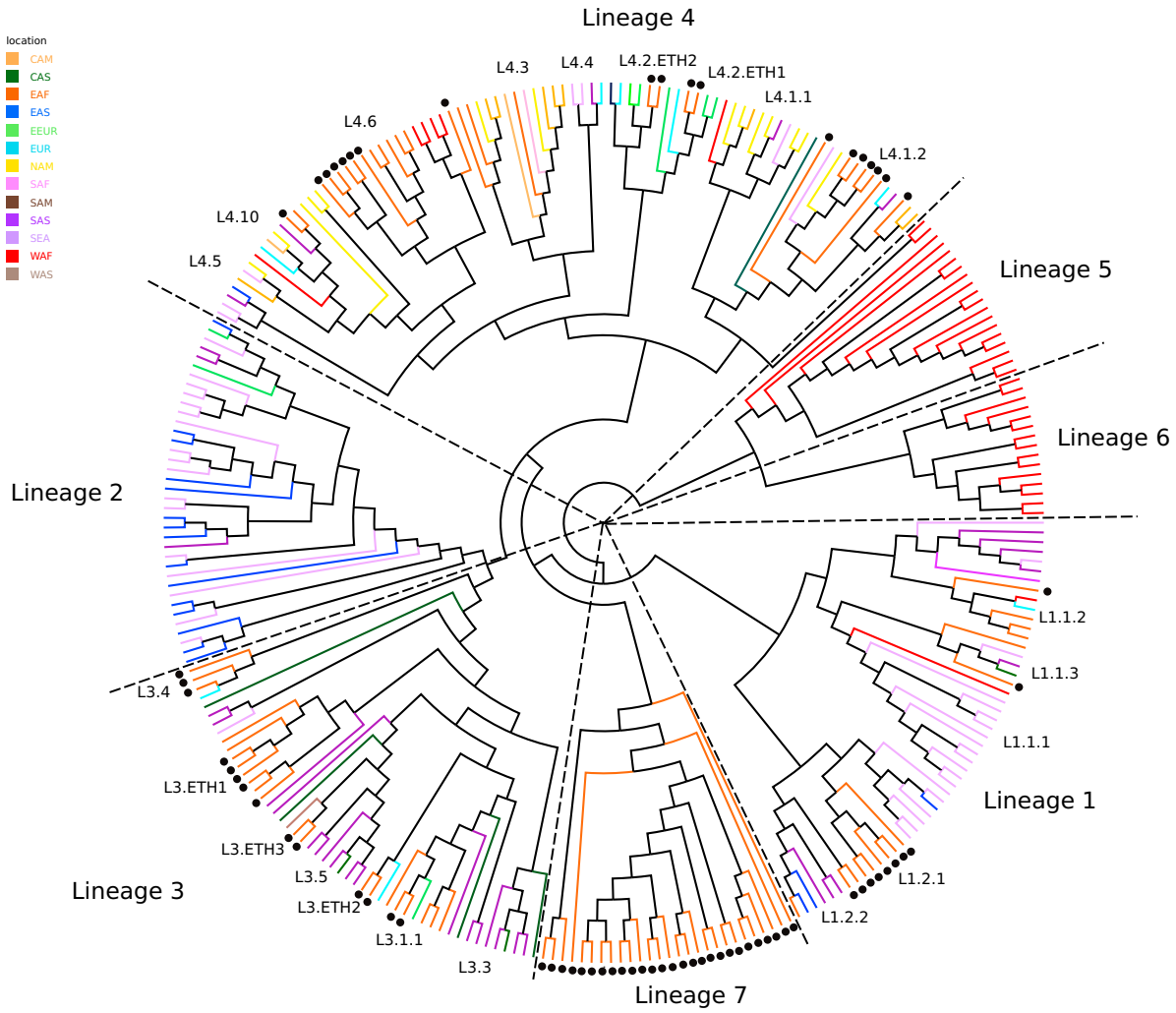


Figure 1. Topology Obtained by Bayesian Analyses as Described in Experimental Procedures

Note that no branch length information is used. The topology is highly congruent with the topology from neighbor-joining analysis (see Figure S1). For both analyses, bootstrap values and Bayesian posterior probability values were higher than 95% in almost all nodes. The tree is rooted with *Mycobacterium africanum* strains that are the most basal clades within the *Mycobacterium tuberculosis* complex [3]. External branches are color coded according to the geographic region of the patient from which the isolate was collected. Black dots indicate that the strain was isolated in Ethiopia. The groups identified within each lineage correspond to the groups delineated using a set of diagnostic SNPs as explained in Experimental Procedures. Groups with no diagnostic SNP or that do not form a monophyletic group within the sub-lineage were labeled LX.ETHX. Geographic region: CAM, Central America; CAS, Central Asia; EAF, East Africa; EAS, East Asia; EEUR, Eastern Europe; EUR, Europe; NAM, North America; SAF, South Africa; SAM, South America; SAS, South Asia; SEA, Southeast Asia; WAF, West Africa; WAS, West Asia.

provisionally name as L3.ETH1, L3.ETH2, and L3.ETH3. Within L1, we assigned Ethiopian strains to four of the eight defined sub-lineages.

We then explored whether the sub-lineages observed in Ethiopia had a likely African origin by comparing them with our global reference dataset [3]. By combining principal-component analysis (PCA) and the sub-lineage classification described above, we classified the groups assigned to the Ethiopian strains as being of likely "African" or "non-African" origin. Figure 2A shows a PCA of the L4 sub-lineages color coded according to their most likely geographic origin. The classic description of L4 as the "Euro-American" lineage is reflected in the high percentage of non-African strains in Figure 2A, though L4 strains are in fact geographically widespread, including genotypes such as sub-lineage L4.6 that are found only in Africa. Within the more cosmopolitan L4 sub-lineages, Ethiopian L4.2 strains cluster with Eurasian strains belonging to the Ural family at two branchpoints in Figure 1, suggesting their introduction into Ethiopia from Central Asia or from some common ancestral homeland [16]. The coalescent point for sub-lineage 4.2 indicates more recent dissemination as compared to the deep-rooted African sub-lineage L4.6 (Figure 1). Similarly, deep-rooted African strains can be found within L3 and L1 (Figures S2 and S3). L3.ETH1 is comprised exclusively of Ethiopian isolates, for example, and deep-rooted L1.2.1 suggests an early introduction into Ethiopia.

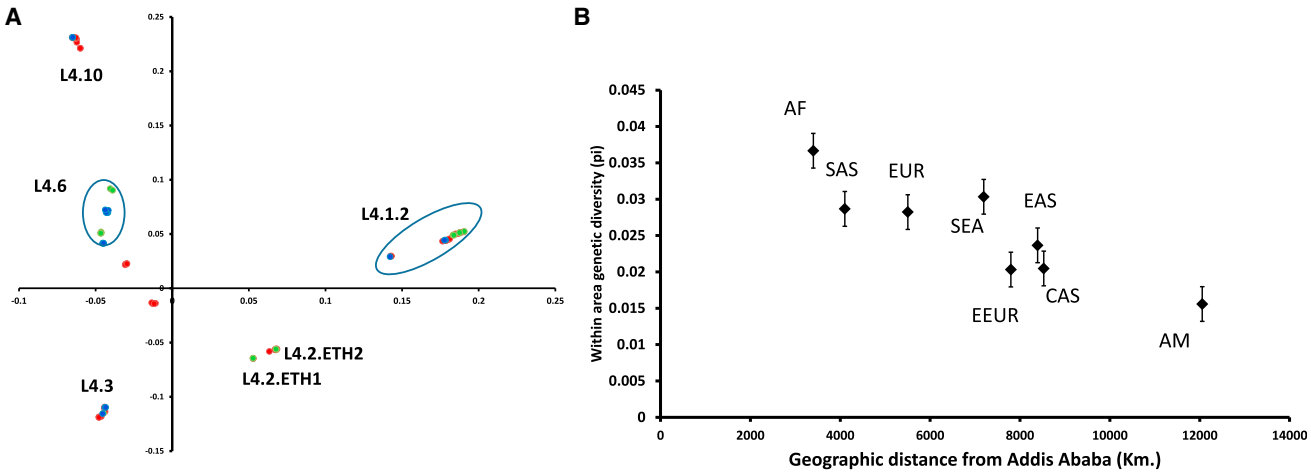


Figure 2. Principal-Component Analysis Using the SNP Matrices Derived from Whole-Genome Analysis

(A) The PCA shown is for lineage 4; the corresponding PCA for lineages 3 and 1 are shown in Figures S2 and S3, respectively. The colors represent strains with known African origin (blue) or Eurasian and American origin (red) from a global reference collection and strains with Ethiopian origin (green).

(B) The correlation between genetic diversity within a geographic area and the geographic distance from Addis Ababa. Error bars indicate the variance in diversity indices within a region. Geographic region: AF, Africa; SAS, South Asia; EUR, Europe; SEA, Southeast Asia; EEUR, Eastern Europe; EAS, East Asia; CAS, Central Asia; AM, America.

In summary, we find strains belonging to sub-lineages with a global distribution as well as strains from sub-lineages specific to Africa or even to Ethiopia. This diversity of *M. tuberculosis* mirrors the complex admixture of African and non-African haplotypes revealed by analysis of human genetic diversity in Ethiopian populations [17, 18].

The Horn of Africa as the Likely Place of Origin of TB

By integrating genome data with our previously reported spoligotype frequencies and collection sites [9], we built contour maps representing the geographic distribution of the principal *M. tuberculosis* genotypes across Ethiopia (Figure 3). L4 dominates across the entire country, L3 is widespread but more frequent in the north, and L1 occurs only in southern Ethiopia. L7 is largely restricted to the highlands around Woldiya and Gondar in northern Ethiopia. The prevalence of L3 in northern Ethiopia is mirrored by the predominant SIT25 spoligotype (L3.ETH1) in neighboring Sudan [19]. Similarly in the south, L1 strains are common in neighboring Somalia [10] (Figure S4). The geographic restriction of L7 could have arisen if the infection was maintained within a stable and isolated human population. Alternatively, it is possible that strains from this lineage have acquired mutations that enhance their infection and transmission in the context of some particular host genetic background but incur a fitness cost in other populations. Similar patterns have been observed in other *M. tuberculosis* populations [20] and in *Helicobacter pylori* [21].

Placing the Ethiopian strains in the context of the global diversity dataset, we plotted within-area genetic diversity against the geographic distances from a midpoint in the African continent and found a strong negative correlation ($r = -0.81$, $p < 0.05$) (Figure 2B). When we repeated the calculation specifying West Africa as the starting point of diversification, we found a similar correlation ($r = -0.79$, $p < 0.05$). However, when we specified the coordinates of Addis Ababa—the capital of Ethiopia—as the point of origin, the correlation became stronger and with higher statistical significance ($r = -0.84$, $p < 0.01$), pointing to East Africa as the likely place of origin of the MTBC. Taken together, our results indicate that the geographic distance to Africa can explain up to 71% of the global genetic diversity of the human-adapted MTBC, which is consistent with serial bottlenecks following initial emergence from a genetically diverse pool in Africa [5]. The strong parallels with patterns identified in human populations [22] and for pathogens such as *Plasmodium falciparum* and *Helicobacter pylori* [23, 24] are striking but consistent with epidemiological observations of robust associations between MTBC genotypes with human populations despite frequent redistribution following increased globalization [20].

Molecular Dating Contradicts the "Virgin Soil" Hypothesis

To review the "virgin soil" hypothesis in the context of an African origin for TB, we tested whether the MTBC lineages and sub-lineages currently circulating in Ethiopia appeared before or after Ethiopia experienced major contacts with Europeans. Although Ethiopia escaped colonization sensu stricto, human genetic diversity data as well as historical records indicate that the region was a bridge between Africa and Eurasia [17, 18]. We analyzed the Ethiopian MTBC genomes in the context of our global diversity dataset and the two competing models for the timing of early events in the evolution of the MTBC. We identified two coalescent points for each of the clusters containing Ethiopian isolates, corresponding to (1) the point of origin of the common ancestor of the cluster and (2) the point at which Ethiopian genotypes diverged from genetically related strains outside of East Africa. We performed BEAST analysis using the 70-thousand-year time frame (MTBC-70 [3]) and the 6-thousand-year time frame (MTBC-6 [25]). The predicted coalescent times obtained using the two models are shown for the major Ethiopian genotypes

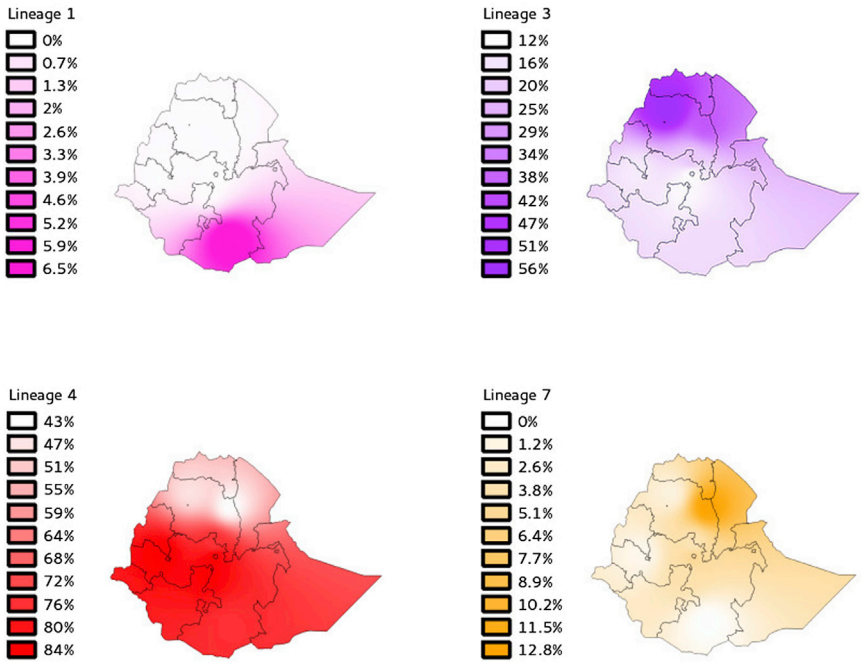


Figure 3. Contour Maps Derived from Point Estimation of the Frequency of Each Lineage in the Different Sampling Locations

The results are based on a previously published *M. tuberculosis* collection [9]. Note that there are different scales for each lineage, reflecting their maximum and minimum frequency across the country.

in Table S2, while Figure 4 shows the estimated coalescent times for the MTBC-6 model.

We were unable to determine which of the two competing models was more likely in the face of historical or pre-historical records (although see Supplemental Experimental Procedures for an extended discussion of this topic). However, both models are clearly incompatible with the view that TB did not exist in Ethiopia or in Sub-Saharan Africa prior to European contact as envisaged in the "virgin soil" hypothesis. Specifically, MTBC-70 dating suggests that the main lineages and sub-lineages were already established 4,000 years ago or earlier (Table S2). The coalescent point for entry of the dominant Ethiopian L4.2.ETH1 cluster is consistent with its arrival in association with the major 3,000-years-ago north-south human migration and origin of the Ethiosemitic languages (95% highest posterior density [HPD] interval: 2,055–3,835 years ago for the divergence within Ethiopia) [26]. Other sub-lineages were already established 5,000 years ago and may be linked to the recent discovery of significant Eurasian admixture in Ethiopia, explained by back migrations of earlier Neolithic farmers probably from Anatolia around that time [18]. In contrast, the MTBC-6 model suggests that some sub-lineages like L4.2.ETH1 (95% HPD interval: 167–313 years ago) or L4.1.2 (95% HPD interval: 404–539 years ago) could have been introduced at the time of early Portuguese contact with Ethiopia in the 16th century or later, while other sub-lineages are several centuries older (Figure 4). In agreement with the PCA plots (Figures 2A, S3, and S4), MTBC-6 predicts that some sub-lineages of lineage 1, lineage 3, and lineage 4 were present in East Africa prior to the 10th century (Table S2; Figure 4), long before the onset of European colonization. The most extreme case of this pattern is L7, with the MTBC-70 time frame suggesting that L7 was branching off around the time of initial human migrations out of Africa (Table S2). MTBC-6 predicts a common ancestor reaching back to the 11th century (95% HPD interval: 806–1,191 years ago) and divergence of L7 from related MTBC strains as early as the third millennium BCE (95% HPD interval: 4,099–4,850 years ago; Table S2; Figure 4). The dating analyses reveal that the same observation of pre-colonial presence of lineages also applies to other parts of Africa. For example, the West Africa-restricted lineages 5 and 6 (L5 and L6), also known as *Mycobacterium africanum*, are predicted to have a common ancestor 5,000 years ago under the MTBC-6 model. In summary, both dating models are incompatible with the "virgin soil" hypothesis for Sub-Saharan Africa.

Conclusions

Understanding the factors underlying the historical persistence or replacement of particular sub-populations of *M. tuberculosis* has potential relevance for predicting future trends in disease epidemiology. There is current evidence that strains belonging to ancient African lineages L5 and L6 are slowly being replaced by L4 in West Africa, and that L2 strains (prevalent in East Asia) are expanding in South Africa (reviewed in [14]), for example, and there is an urgent need to identify the mechanisms that determine expansion or restriction of emerging drug-resistant genotypes. Successful introduction of new genotypes could be driven by major population migrations or by fitness properties linked to host genotype or social environment [27]. A simple conceptual model to account for the population structure in Ethiopia would envisage that prolonged stable co-evolution of human and microbial populations favors a relatively benign infection that optimizes mutual survival of both partners, whereas an unstable environment favors expansion of more aggressive microbial genotypes irrespective of their long-term impact on host populations. TB epidemiology in colonial Africa would then reflect replacement of a stable *M. tuberculosis* population with introduction of aggressive genotypes selected in industrialized Europe alongside expansion of opportunist indigenous genotypes with the ability to exploit opportunities of African urbanization [28]. This can explain the high heterogeneity in infection rates across Sub-Saharan Africa observed after colonial contact [29]. Consistent with this model, the most recently introduced sub-lineage L4.1.2 is associated with large transmission clusters in Ethiopia [9]; an independent genotyping study in the Amhara region similarly demonstrated higher transmission and drug resistance associated with spoligotypes corresponding to L4.1.2 (Haarlem) and L4.2.ETH1 (named NW-ETH3 in [30, 31]). In

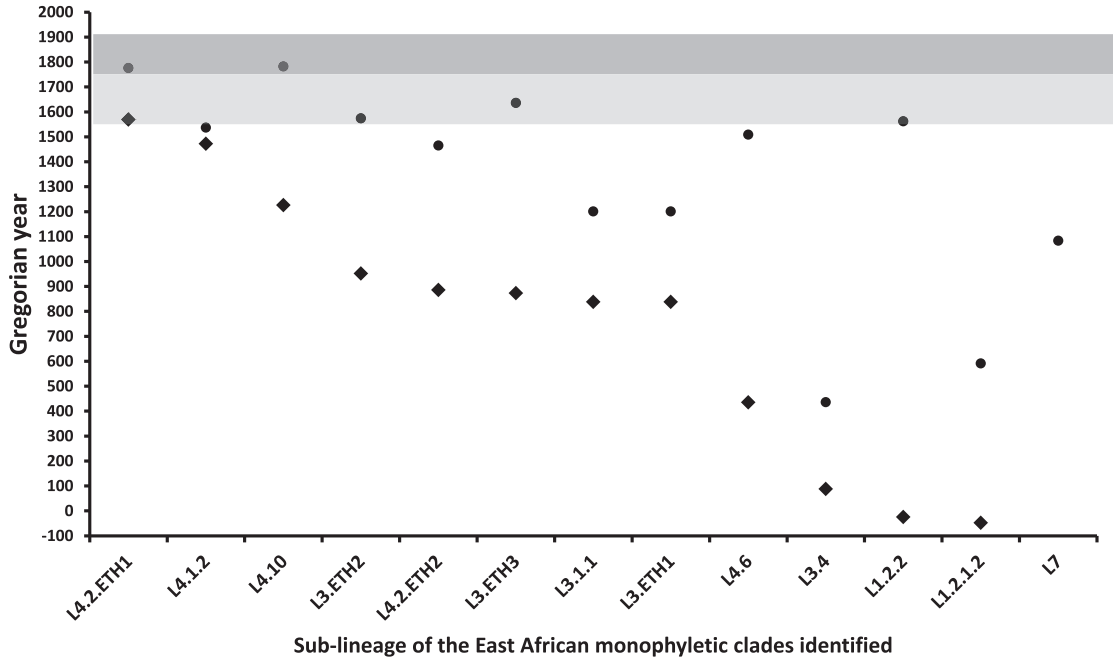


Figure 4. Representation of Dating Events for MTBC Sub-lineages Included in This Study

Dots (●) represents the age of the most recent common ancestor of the different Ethiopian sub-lineages circulating today. Diamonds (◆) represent the split of those Ethiopian groups from the closest non-Ethiopian strains in the global dataset described in [3]. Note that for lineage 7, the split from other MTBC was the third millennium BCE and is not represented. Light gray highlights the time of first sporadic European contacts in Ethiopia (16^{th}–19^{th} centuries CE). Dark gray highlights the time frame for a more continuous contact between Ethiopia and foreign nations (19^{th} century CE and onward). The data presented show only the results for the MTBC-6 model in which the whole complex is predicted to be around 6,000 years old [25]; the results for the MTBC-70 model where all lineages were already established by 4,000 years ago are not shown. (See Table S2.)

contrast, patients infected with L7 strains appear to report later to the health clinics and tend not to be associated with recent transmission and multidrug resistance [30].

EXPERIMENTAL PROCEDURES

Preparation of Genomic DNA and Genome Sequencing

All Ethiopian samples were obtained from a previous study [9]. Selected MTBC strains were cultured from frozen stocks on Middlebrook 7H11 agar. One single colony per strain was then sub-cultured, cells were harvested by heat inactivation, and genomic DNA was purified using a standard protocol [32] and utilized for sequencing on an Illumina HiSeq platform at GATC Biotech (Konstanz, Germany) with coverage ranging between 150 and 244 reads per base pair. Reads were mapped to a reconstructed most recent common ancestor of the MTBC as described previously [33]. For each strain, a combination of BWA mapping and SAMtools SNP calling [34] was used with the parameters set as described earlier [9]. We controlled for false positives by removing calls falling in repetitive regions [33]. SNP calls should have been present in at least 10 reads, with SNP mapping qualities of the bases higher than 20. The 66 sequenced genomes were combined with a recently published whole-genome dataset of 219 global MTBC isolates [3].

Statistical Analyses

To define the population structure of *M. tuberculosis* in Ethiopia, we scanned the strains for diagnostic SNPs defining phylogenetic groups as recently published [15]. To associate the different Ethiopian phylogenetic groups with global African or non-African clades, we carried out PCA using Stata [35]. To explore the possible relationship between genetic diversity and geographic distance, we classified MTBC strains into 13 broad geographic areas. These areas are approximately the same as those defined by the United Nations [3]. The within-area genetic diversity (pi) was calculated using DnaSP [20]. We used the Pearson correlation coefficient to test for correlation between pi for MTBC and the mean geographic distance between the country of origin of the patient and an arbitrary point representing (1) a midpoint in Africa (latitude 6.793981° N, longitude 21.299250° E), (2) Addis Ababa (8.984711° N, 38.754112° E), or (3) West Africa (12.433722° N, 1.102581° W).

Contour maps to infer the frequency of each lineage across the country were built using available genotyping data of almost 1,000 *M. tuberculosis* isolates previously generated [9]. We inferred the incidence of the lineage across the country using the inverse distance weighting interpolation with a power value of 2.7. All analyses were performed using QGIS v2.8, which is freely available at http://www.qgis.org/en/site/.

Phylogenetic and Dating Analyses

For phylogenetic inference, we used a core genome alignment of SNPs after removing columns with gaps (deletions or heterozygous calls). Phylogenetic analyses were performed using MEGA for the neighbor-joining (NJ) tree (Tamura-3 parameters) with 1,000 bootstrap pseudo-replicates [36]. RAxML version 8.0.0 [37] was used for the maximum-likelihood approach specifying a general time-reversible model of nucleotide evolution and estimating five gamma categories and 1,000 bootstrap pseudo-replicates. BEAST v1.7.5 was used for the Bayesian topology inference as specified below [38]. The application of different inference methods or models of nucleotide evolution did not have an impact on the phylogeny, with all relevant clades discussed in this manuscript having a support value of 100 for the three inference methods (see Figure 1 for the maximum-likelihood topology and Figure S1 for details of support values of the main clades discussed).

The dating analyses were performed using similar parameters as in earlier publications [3, 25]. Briefly, the phylogenetic tree was calibrated with the two existing competing models for predicting the origin of the MTBC: a 70,000-year-old ancestor [3] or a 6,000-year-old ancestor [25] was pre-specified in BEAST. We used an uncorrelated log-normal distribution to model the variation of the substitution rate among branches, a Hasegawa-Kishino-Yano

nucleotide substitution model, and a skyline demographic model with ten different intervals. We ran between six and ten chains during 5×10^7 generations and sampled every 1×10^4 generations to assure independent convergence. The burn-in was set to 20%. Convergence was evaluated in Tracer [39] with all relevant parameters reaching an effective sample size > 100.

ACCESSION NUMBERS

Raw FASTQ files have been deposited at the EBI Sequence Read Archive under accession numbers SRA: PRJEB9201, PRJEB3163, PRJEB3124.

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, two tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2015.10.061.

AUTHOR CONTRIBUTIONS

Study design was performed by S. Berg, I.C., S.G., D.Y., A.A., E.H., R.G.H., B.D.R., R.T., B.G., and G.A. Data collection (sampling, mycobacterial culturing, molecular typing) was performed by T.K., E.H., S. Bekele, W.M., B.G., I.C., and S. Berg. Data analyses were performed by I.C., S. Berg, D.Y., S.G., G.A.G., D.S., and A.A. The manuscript was written primarily by I.C., S. Berg, D.Y., S.G., and A.A. All authors read the manuscript and approved its publication. I.C. and S. Berg contributed equally and jointly directed the work.

ACKNOWLEDGMENTS

We thank collaborating health facilities in Ethiopia and Armauer Hansen Research Institute (AHRI) staff, particularly Rebuma Firdessa, Araya Mengistu, and Endalamaw Gadisa for various assistance. AHRI receives core support from Sida (Sweden) and NORAD (Norway). This study was sponsored by the Wellcome Trust (grant number 075833/A/04/Z) under their "Animal Health in the Developing World" initiative. I.C. and G.A.G. were supported by Ramón y Cajal Spanish research grant RYC-2012-10627, MINECO research grant SAF2013-43521-R, and the European Research Council (ERC) (638553-TB-ACCELERATE). D.S. and S.G. were supported by the Swiss National Science Foundation (PP00P3_150750) and ERC (309540-EVODRTB). D.S. was supported by the Swiss HIV Cohort study (grant 740). This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK, the Medical Research Council UK, and the Wellcome Trust.

Received: September 11, 2015
Revised: October 26, 2015
Accepted: October 28, 2015
Published: December 10, 2015

REFERENCES

1. Collins, T.F.B. (1982). The history of southern Africa's first tuberculosis epidemic. S. Afr. Med. J. *62*, 780–788.
2. Cummins, S.L. (1929). "Virgin Soil"—and after. A working conception of tuberculosis in children, adolescents, and aborigines. BMJ *2*, 39–41.
3. Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., et al. (2013). Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat. Genet. *45*, 1176–1182.
4. Supply, P., Marceau, M., Mangenot, S., Roche, D., Rouanet, C., Khanna, V., Majlessi, L., Criscuolo, A., Tap, J., Pawlik, A., et al. (2013). Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. Nat. Genet. *45*, 172–179.
5. Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A., Feldman, M.W., and Gagneux, S. (2008). High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol. *6*, e311.
6. Takiff, H.E., and Feo, O. (2015). Clinical value of whole-genome sequencing of *Mycobacterium tuberculosis*. Lancet Infect. Dis. *15*, 1077–1090.
7. Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., et al. (2002). A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc. Natl. Acad. Sci. USA *99*, 3684–3689.
8. Boritsch, E.C., Supply, P., Honoré, N., Seemann, T., Stinear, T.P., and Brosch, R. (2014). A glimpse into the past and predictions for the future: the molecular evolution of the tuberculosis agent. Mol. Microbiol. *93*, 835–852.
9. Firdessa, R., Berg, S., Hailu, E., Schelling, E., Gumi, B., Erenso, G., Gadisa, E., Kiros, T., Habtamu, M., Hussein, J., et al. (2013). Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. Emerg. Infect. Dis. *19*, 460–463.
10. Blouin, Y., Hauck, Y., Soler, C., Fabre, M., Vong, R., Dehan, C., Cazajous, G., Massoure, P.-L., Kraemer, P., Jenkins, A., et al. (2012). Significance of the identification in the Horn of Africa of an exceptionally deep branching *Mycobacterium tuberculosis* clade. PLoS ONE *7*, e52841.
11. Blouin, Y., Cazajous, G., Dehan, C., Soler, C., Vong, R., Hassan, M.O., Hauck, Y., Boulais, C., Andriamanantena, D., Martinaud, C., et al. (2014). Progenitor "*Mycobacterium canettii*" clone responsible for lymph node tuberculosis epidemic, Djibouti. Emerg. Infect. Dis. *20*, 21–28.
12. Pankhurst, R. (1990). The Medical History of Ethiopia (The Red Sea Press).
13. World Health Organization (2014). Global Tuberculosis Report 2014 (WHO).
14. Coscolla, M., and Gagneux, S. (2014). Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin. Immunol. *26*, 431–444.
15. Coll, F., McNerney, R., Guerra-Assunção, J.A., Glynn, J.R., Perdigão, J., Viveiros, M., Portugal, I., Pain, A., Martin, N., and Clark, T.G. (2014). A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat. Commun. *5*, 4812.
16. Mokrousov, I. (2012). The quiet and controversial: Ural family of *Mycobacterium tuberculosis*. Infect. Genet. Evol. *12*, 619–629.
17. Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J., Geberhiwot, T., Usanga, E., and Villems, R. (2004). Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am. J. Hum. Genet. *75*, 752–770.
18. Llorente, M.G., Jones, E.R., Eriksson, A., Siska, V., Arthur, K.W., Arthur, J.W., Curtis, M.C., Stock, J.T., Coltorti, M., Pieruccini, P., et al. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science. Published online October 8, 2015. http://dx.doi.org/10.1126/science.aad2879.
19. Sharaf Eldin, G.S., Fadl-Elmula, I., Ali, M.S., Ali, A.B., Salih, A.L., Mallard, K., Bottomley, C., and Mcnerney, R. (2011). Tuberculosis in Sudan: a study of *Mycobacterium tuberculosis* strain genotype and susceptibility to anti-tuberculosis drugs. BMC Infect. Dis. *11*, 219.
20. Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., et al. (2006). Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. USA *103*, 2869–2873.
21. Kodaman, N., Pazos, A., Schneider, B.G., Piazuelo, M.B., Mera, R., Sobota, R.S., Sicinschi, L.A., Shaffer, C.L., Romero-Gallo, J., de Sablet, T., et al. (2014). Human and *Helicobacter pylori* coevolution shapes the risk of gastric disease. Proc. Natl. Acad. Sci. USA *111*, 1455–1460.
22. Henn, B.M., Cavalli-Sforza, L.L., and Feldman, M.W. (2012). The great human expansion. Proc. Natl. Acad. Sci. USA *109*, 17758–17764.
23. Tanabe, K., Mita, T., Jombart, T., Eriksson, A., Horibe, S., Palacpac, N., Ranford-Cartwright, L., Sawai, H., Sakihama, N., Ohmae, H., et al. (2010). *Plasmodium falciparum* accompanied the human expansion out of Africa. Curr. Biol. *20*, 1283–1289.
24. Linz, B., Balloux, F., Moodley, Y., Manica, A., Liu, H., Roumagnac, P., Falush, D., Stamer, C., Prugnolle, F., van der Merwe, S.W., et al. (2007).

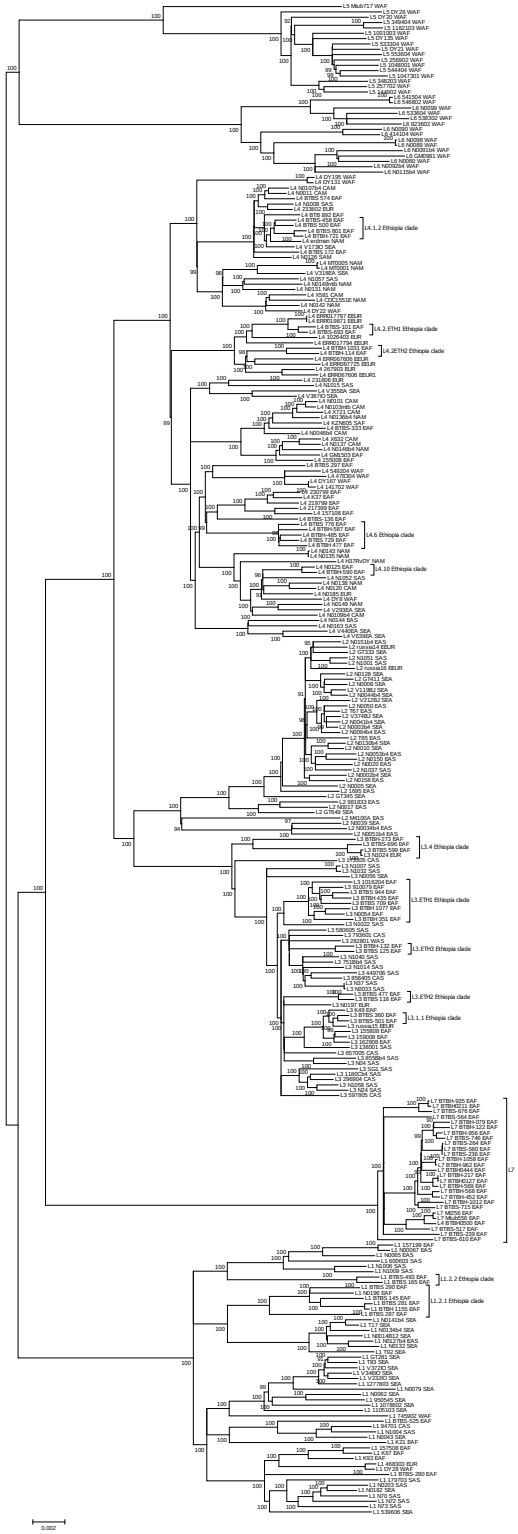
An African origin for the intimate association between humans and *Helicobacter pylori*. Nature *445*, 915–918.

25. Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S.A., Bryant, J.M., Harris, S.R., Schuenemann, V.J., et al. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature *514*, 494–497.
26. Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., et al. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am. J. Hum. Genet. *91*, 83–96.
27. Comas, I., and Gagneux, S. (2009). The past and future of tuberculosis research. PLoS Pathog. *5*, e1000600.
28. Comas, I., and Gagneux, S. (2011). A role for systems epidemiology in tuberculosis research. Trends Microbiol. *19*, 492–500.
29. Cummins, S.L. (1935). Studies of tuberculosis among African natives: reports to the Medical Research Council. Tubercle *1935*, 7–15.
30. Yimer, S.A., Norheim, G., Namouchi, A., Zegeye, E.D., Kinander, W., Tønjum, T., Bekele, S., Mannsåker, T., Bjune, G., Aseffa, A., and Holm-Hansen, C. (2015). *Mycobacterium tuberculosis* lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in Amhara Region, Ethiopia. J. Clin. Microbiol. *53*, 1301–1309.
31. Agonafir, M., Lemma, E., Wolde-Meskel, D., Goshu, S., Santhanam, A., Girmachew, F., Demissie, D., Getahun, M., Gebeyehu, M., and van Soolingen, D. (2010). Phenotypic and genotypic analysis of multidrug-resistant tuberculosis in Ethiopia. Int. J. Tuberc. Lung Dis. *14*, 1259–1265.
32. van Soolingen, D., Hermans, P.W.M., de Haas, P.E.W., Soll, D.R., and van Embden, J.D. (1991). Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J. Clin. Microbiol. *29*, 2578–2586.
33. Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., Galagan, J., Niemann, S., and Gagneux, S. (2012). Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat. Genet. *44*, 106–110.
34. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
35. StataCorp (2007). Stata Statistical Software: Release 10 (StataCorp).
36. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731–2739.
37. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.
38. Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. *7*, 214.
39. Rambaut, A., Suchard, M.A., Xie, D., and Drummond, A.J. (2014). Tracer v1.6. http://beast.bio.ed.ac.uk/Tracer.

Current Biology
Supplemental Information

Population Genomics of *Mycobacterium tuberculosis* in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa

Iñaki Comas, Elena Hailu, Teklu Kiros, Shiferaw Bekele, Wondale Mekonnen, Balako Gumi, Rea Tschopp, Gobena Ameni, R. Glyn Hewinson, Brian D. Robertson, Galo A. Goig, David Stucki, Sebastien Gagneux, Abraham Aseffa, Douglas Young, and Stefan Berg



Ethiopian clade	NJ	BS	ML
L4.1.2	100	1	100
L4.2.ETH2	100	1	100
L4.2.ETH1	100	1	100
L4.10	100	1	100
L4.6	100	1	100
L3.1.1	100	1	100
L3.4	100	1	100
L3.ETH3	100	1	100
L3.ETH2	100	1	100
L3.ETH1	100	1	100
L1.2.1	100	1	100
L1.2.2	100	1	100
L7	100	1	100

Figure S1. Maximum Likelihood phylogeny under the General-Time-Reversible model of nucleotide evolution (five gamma categories; 1,000 bootstrap pseudo-replicates). The scale is proportional to the number of substitutions per polymorphic site. Similar topologies were obtained using Neighbour-Joining and Bayesian phylogenetic inference approaches. Insert table shows the support values for the three approaches of the main clades discussed in the text (NJ, Neigbour-Joining bootstrap; BS, Bayesian Support; ML, Maximum-Likelihood bootstrap). Related to Figure 1 and Figure 4.

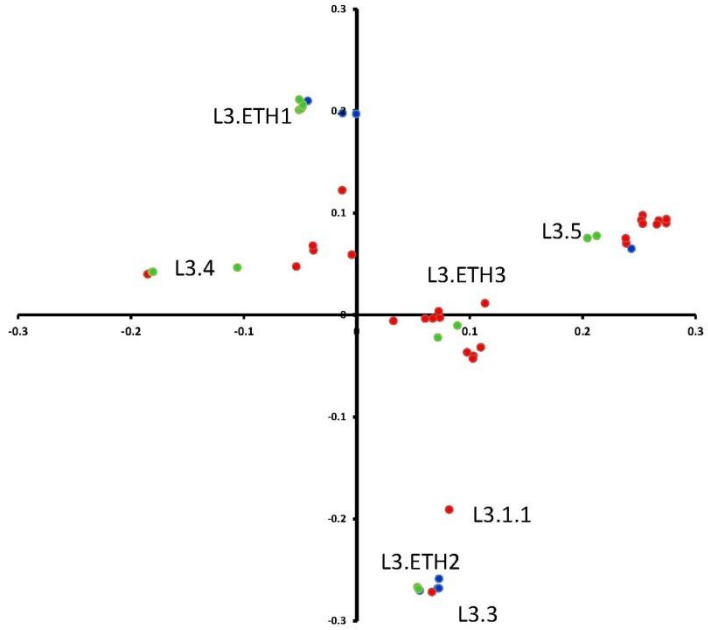


Figure S2. Principal component analysis for the Lineage 3 sub-lineages present in Ethiopia in relation with a global reference strain. The colors represent strains with known African origin (blue), Ethiopian origin (green), or Eurasian and American origin (red). Related to Figure 2.

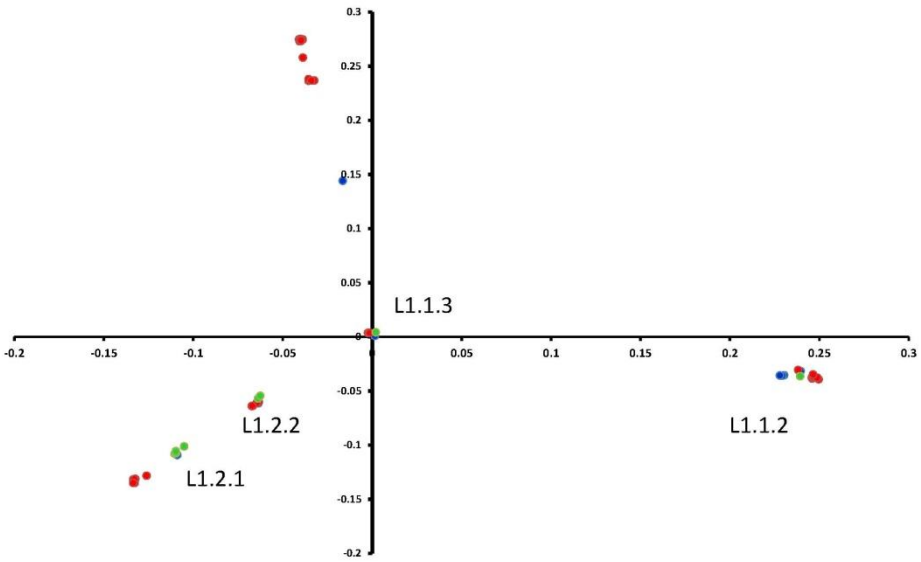


Figure S3. Principal component analysis for the Lineage 1 sub-lineages present in Ethiopia in relation with a global reference strain. The colors represent strains with known African origin (blue), Ethiopian origin (green), or Eurasian and American origin (red). Related to Figure 2.

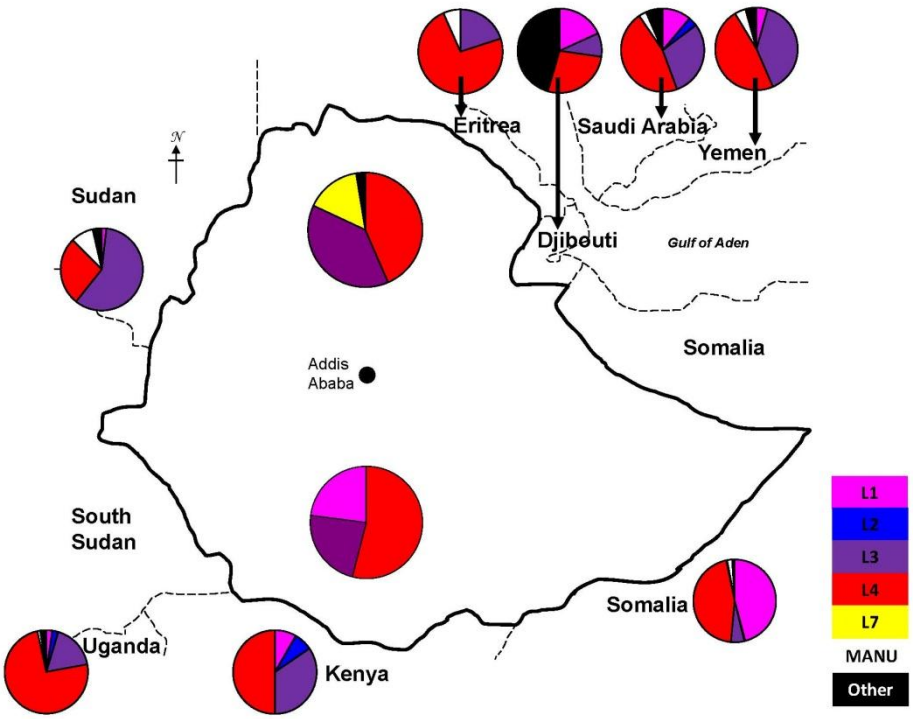


Figure S4. Geographic distribution of main MTBC lineages in the North and the South of Ethiopia based on genotyping of nearly 1000 MTBC isolates [S1]. Corresponding frequencies of MTBC lineages in the neighboring countries have been estimated based on spoligotype data of MTBC isolates registered at the SITVIT database [S2]. The records extracted from this database may not reflect the true prevalence of different lineages in these countries. "MANU" refers to the MANU family [S2] while "Others" refers to isolates of *Mycobacterium canettii* or unknown MTBC lineage. Related to Figure 3.

Table S1. Summary of information of *Mycobacterium tuberculosis* strains from Ethiopia analysed by genome sequencing in this study. Related to Figure 1.

Lineage	Classification	Strain ID	Coverage	Disease type	Collection site	Patient ethnicity	Spoligotype	**Spoligotype pattern** (filled cell = spacer present; empty cell = spacer absent)																																										
SNP based	Coll et al. 2015		mean				SITVIT2 DB	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
L1	L1.2.1	BTBS0145	243.85	Pulmonary TB	Negelle	Oromo	SIT 43	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■					■		■	■	■	■	■	■	■	■	■
L1	L1.2.2	BTBS0165	196.146	Pulmonary TB	Negelle	Oromo	SIT 48	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■		■	■	■	■	■		■	■	■
L1	L1.1.2	BTBS0280	235.585	Pulmonary TB	Filtu	Somali	SIT 10	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	■	■	■	■	■					■		■	■	■	■	■	■	■	■	■
L1	L1.2.1	BTBS0281	233.046	Pulmonary TB	Filtu	Somali	SIT 204	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■		■	■	■	■	■	■	■	■	■
L1	L1.2.1	BTBS0287	233.598	Pulmonary TB	Negelle	Oromo	SIT 591	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■					■		■	■	■	■	■	■	■	■	■
L1	L1.2.1	BTBS0290	215.48	Pulmonary TB	Negelle	Oromo	SIT 3332	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■						■					■		■		■	■	■	■	■	■	■
L1	L1.2.2	BTBS0493	237.682	Pulmonary TB	Negelle	Oromo	SIT 48	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■		■	■	■	■	■		■	■	■
L1	L1.1.3	BTBS0525	235.608	Pulmonary TB	Negelle	Oromo	SIT 924	■	■	■	■	■	■	■	■	■	■	■														■	■	■	■					■		■	■		■	■	■	■	■	■
L1	L1.2.1	BTBH1155	239.488	TB lymphadenitis	Filtu	Somali	SIT 3335	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■					■		■	■	■	■	■	■	■	■	■
L3	L3.ETH2	BTBS0118	196.109	Pulmonary TB	Fiche	Oromo	SIT 26	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■													■	■	■	■	■	■	■	■	■
L3	L3.ETH3	BTBS0125	232.262	Pulmonary TB	Negelle	Oromo	SIT 247	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■													■	■	■			■	■	■	■
L3	L3.ETH3	BTBH0132	212.985	TB lymphadenitis	Woldiya	Amhara	Orphan	■							■	■	■	■	■	■	■	■	■	■	■	■	■	■	■																					
L3	L3.4	BTBH0273	236.887	TB lymphadenitis	Woldiya	Amhara	SIT 1200	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
L3	L3.ETH1	BTBH0351	236.904	TB lymphadenitis	Woldiya	Amhara	SIT 25	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■													■	■			■	■	■	■	■
L3	L3.1.1	BTBS0360	223.748	Pulmonary TB	Gondar	Oromo	SIT 21	■	■	■					■	■		■	■	■	■	■	■	■	■	■																		■	■	■	■	■	■	■	■
L3	L3.ETH1	BTBH0435	228.314	TB lymphadenitis	Butajira	Gurage	SIT 25	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■													■	■			■	■	■	■	■
L3	L3.ETH2	BTBS0477	229.68	Pulmonary TB	Negelle	Oromo	SIT 26	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■													■	■	■	■	■	■	■	■	■
L3	L3.1.1	BTBS0501	231.567	Pulmonary TB	Negelle	Oromo	SIT 21	■	■	■					■	■		■	■	■	■	■	■	■	■	■																		■	■	■	■	■	■	■	■
L3	L3.4	BTBS0599	230.005	Pulmonary TB	Butajira	Amhara	SIT 1																																			■	■	■	■	■	■	■	■	■
L3	L3.4	BTBS0696	216.589	Pulmonary TB	Negelle	Amhara	SIT 1																																			■	■	■	■	■	■	■	■	■
L3	L3.ETH1	BTBS0709	244.302	Pulmonary TB	Fiche	Amhara	SIT 25	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■													■	■			■	■	■	■	■
L3	L3.ETH1	BTBS0944	223.858	Pulmonary TB	Gondar	Amhara	SIT 25	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■													■	■			■	■	■	■	■
L3	L3.ETH1	BTBH1077	240.882	TB lymphadenitis	Woldiya	Amhara	SIT 25	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■	■	■	■													■	■			■	■	■	■	■
L4	L4.2.ETH1	BTBS0101	238.678	Pulmonary TB	Gondar	Oromo	SIT 149	■	■	■	■	■	■	■	■	■											■	■	■	■	■	■	■	■	■	■	■	■	■					■	■	■	■	■	■	■
L4	L4.2.ETH2	BTBH0114	237.376	TB lymphadenitis	Butajira	Silti	SIT 3135	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■																			
L4	L4.6	BTBS0136	235.904	Pulmonary TB	Filtu	Somali	SIT 584	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	■	■		■	■	■
L4	L4.6	BTBS0297	218.856	Pulmonary TB	Negelle	Oromo	SIT 2793	■																		■	■	■	■		■	■	■	■	■	■	■	■	■					■	■	■	■	■	■	■
L4	L4.3	BTBS0333	229.551	Pulmonary TB	Negelle	Oromo	SIT 4																									■	■	■	■	■	■	■	■					■	■	■	■	■	■	■
L4	L4.1.2	BTBS0458	231.67	Pulmonary TB	Gondar	Amhara	SIT 3313																									■							■					■	■			■	■	■
L4	L4.6	BTBH0477	225.319	TB lymphadenitis	Ghimbi	Oromo	SIT 3337	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■					■	■	■	■					■	■	■	■	■	■	■
L4	L4.6	BTBH0485	235.216	TB lymphadenitis	Ghimbi	Amhara	SIT 3137	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	■	■	■	■	■	■
L4	L4.1.2	BTBS0500	200.73	Pulmonary TB	Negelle	Oromo	SIT 134	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■					■	■			■	■	■
L4	L4.1.2	BTBS0574	223.594	Pulmonary TB	Fiche	Oromo	SIT 50	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■					■	■	■	■	■	■	■
L4	L4.6	BTBH0587	209.579	TB lymphadenitis	Gondar	Oromo	SIT 37	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	■	■	■	■	■	■
L4	L4.10	BTBH0590	209.258	TB lymphadenitis	Ghimbi	Oromo	SIT 53	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	■	■	■	■	■	■
L4	L4.2.ETH1	BTBS0693	232.159	Pulmonary TB	Negelle	Oromo	SIT 149	■	■	■	■	■	■	■	■	■											■	■	■	■	■	■	■	■	■	■	■	■	■					■	■	■	■	■	■	■
L4	L4.1.2	BTBH0721	225.859	TB lymphadenitis	Butajira	Gurage	SIT 3134	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■		■					■	■	■	■	■	■	■
L4	L4.6	BTBS0729	218.798	Pulmonary TB	Holeta	data missing	SIT3314							■	■		■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					■	■	■	■	■	■	■
L4	L4.6	BTBS0776	193.269	Pulmonary TB	Ghimbi	Oromo	Orphan	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	■	■	■	■	■	■	■																			
L4	L4.1.2	BTBS0801	244.19	Pulmonary TB	Butajira	data missing	SIT 3134	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■		■					■	■	■	■	■	■	■
L4	L4.2.ETH2	BTBH1031	218.033	TB lymphadenitis	Woldiya	Amhara	SIT 46	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■																			
L7	Not classified	BTBH0079	240.671	TB lymphadenitis	Woldiya	Amhara	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0122	237.462	TB lymphadenitis	Woldiya	Amhara	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0127	190.361	TB lymphadenitis	Woldiya	Amhara	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0211	198.175	TB lymphadenitis	Woldiya	Gurage	SIT 3336	■	■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	
L7	Not classified	BTBH0217	241.798	TB lymphadenitis	Butajira	Silti	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBS0236	235.858	Pulmonary TB	Fiche	Oromo	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBS0239	238.313	Pulmonary TB	Fiche	Oromo	SIT 910	■	■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBS0264	238.319	Pulmonary TB	Fiche	Oromo	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0444	185.465	TB lymphadenitis	Woldiya	Amhara	Orphan	■	■	■																						■							■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0452	190.352	TB lymphadenitis	Woldiya	Tigre	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0500	150.196	TB lymphadenitis	Woldiya	Amhara	SIT 910	■	■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBS0517	232.237	Pulmonary TB	Woldiya	Amhara	SIT 910	■	■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBS0564	237.648	Pulmonary TB	Fiche	data missing	SIT 910	■	■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0568	236.93	TB lymphadenitis	Woldiya	Amhara	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0569	241.15	TB lymphadenitis	Woldiya	Amhara	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBS0580	233.871	Pulmonary TB	Fiche	Oromo	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBS0610	238.871	Pulmonary TB	Butajira	data missing	SIT 910	■	■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBS0676	235.172	Pulmonary TB	Addis Ababa	data missing	SIT 910	■	■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0715	228.811	TB lymphadenitis	Fiche	Amhara	Orphan		■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBS0746	216.68	Pulmonary TB	Woldiya	Amhara	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0935	221.315	TB lymphadenitis	Butajira	Gurage	SIT 3336	■	■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	
L7	Not classified	BTBH0956	210.431	TB lymphadenitis	Woldiya	Amhara	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH0962	228.625	TB lymphadenitis	Woldiya	data missing	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH1012	219.632	TB lymphadenitis	Butajira	data missing	SIT 910	■	■	■																						■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■
L7	Not classified	BTBH1058	238.187	TB lymphadenitis	Woldiya	Amhara	SIT 1729	■	■	■																						■					■	■	■	■	■	■	■	■	■	■	■	■	■	■

Table S2. Coalescent times in MTBC evolution for the most frequent Ethiopian clades identified in this study. Median height values for two coalescent events are shown for respective sub-lineage (see text for details). The dates are given in years before present. Related to Figure 4.

	MTBC-70 (years ago)	**MTBC-6 (years ago)**	**MTBC-70 (years ago)**	**MTBC-6 (years ago)**
Coalescent event	**Split from closest non-East African**		**Divergence within East Africa**	
L4.1.2	6,565	538	5,759	473
L4.2.ETH2	13,849	1,124	6,628	545
L4.2.ETH1	5,548	441	2,863	234
L4.10	9,904	784	2,847	228
L4.6	19,669	1,575	6,009	501
L3.1.1	14,300	1,172	9,831	809
L3.4	23,763	1,922	19,383	1,574
L3.ETH3	15,956	1,137	5,201	374
L3.ETH2	14,107	1,058	4,392	436
L3.ETH1	10,249	1,172	7,215	809
L1.2.1	26,792	2,058	18,086	1,419
L1.2.2	26,567	2,035	5,484	448
L7	58,276	4,461	11,280	927

Supplemental Experimental Procedures

Correlation between dating analyses and known Ethiopia historical events

Based on the MTBC-70 model, the dominant Ethiopian sub-lineage L4.2 (15% of all 950 isolates based on spoligotype analysis [S1]) has a coalescent date consistent with its arrival in association with the major north-south human migration and origin of the Ethiosemitic languages around 3 thousand years ago [S3] (Table S2). L4.2 includes the "Ural" spoligotype family commonly found in the Pontic region north of the Black Sea and in Iran [S4], and a plausible scenario could involve infection of a Neolithic human population in the Levant by the L4.2 progenitor, followed by divergent north/south migrations. Sub-lineage L4.6 has an older coalescent point, corresponding to 20 thousand years ago in the MTBC-70 model. A potential link with patterns of human migration is provided by Eurasian mitochondrial DNA haplogroups U6 and M1 which were present in Africa during the Upper Paleolithic period prior to Neolithic expansion [S5]. L4.6 is currently restricted to Africa, and its common occurrence in Uganda, Cameroon and neighbouring countries suggests that it could have been disseminated across southern Africa in association with the Bantu expansion between 1000 BC and 500 AD. The remaining two sub-lineages of L4 observed in Ethiopia may represent recent colonial spread from Europe as evidenced, for example, by the global distribution of the spoligotype family known as "Latin American Mediterranean" (LAM) [S6]. Coalescent analysis of L3 suggests a similar pattern of multiple introductions into Ethiopia. Dating by the MTBC-70 model suggests that the major sub-lineage L3.ETH1 was established in Ethiopia prior to the Neolithic period. The geographic distribution of L3.ETH1, together with its early divergence from sub-lineages present in the Indian sub-continent, suggests an introduction into Ethiopia by ancient land-based human migrations. Coalescent analysis of other Ethiopian L3 isolates mapping to sub-lineages with a broader geographic distribution would be consistent with a more recent introduction through trade routes across the Indian Ocean.

Coalescent times estimated according to the MTBC-6 model suggest a much more recent origin and spread of *M. tuberculosis* in Ethiopia between the 15th and 19th centuries (Table S2). These dates are difficult to reconcile with major population migrations into Ethiopia, though historical records show continuous internal migrations with southward movement of Tigrean populations during ancient and medieval periods, northward migration of Oromos between the 16th and early 18th centuries, and southern movement of Amharas in the 19th century. Documentation of a recent expansion of L2 genotype in South Africa demonstrates that large-scale population movements are not a prerequisite for effective dissemination of a novel strain within a susceptible population, however, while the absence of an association of the West African slave trade with Lineages 5 and 6 shows that *M. tuberculosis* genotypes do not inevitably follow population movements. The predicted 16th century introduction of L4.2 into Ethiopia could reflect a period of increasing openness to European adventurers, with L4.2.ETH1 aligning with the timing of a Portuguese military expedition to support the resistance of Ethiopian Christians against Muslim incursions. The 12th century dating of L3.4

and L3.ETH1 would be consistent with the dissemination of these clades by Indian Ocean trade.

Supplemental References

S1. Firdessa, R., Berg, S., Hailu, E., Schelling, E., Gumi, B., Erenso, G., Gadisa, E., Kiros, T., Habtamu, M., Hussein, J., et al. (2013). Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. Emerg. Infect. Dis. *19*, 460–463.

S2. Demay, C., Liens, B., Burguiere, T., Hill, V., Couvin, D., Millet, J., Mokrousov, I., Sola, C., Zozio, T., and Rastogi, N. (2012). SITVITWEB--a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect. Genet. Evol. *12*, 755–766.

S3. Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Romero, I. G., Ayub, Q., Mehdi, S. Q., Thomas, M. G., Luiselli, D., et al. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the ethiopian gene pool. Am. J. Hum. Genet. *91*, 83–96.

S4. Mokrousov, I. (2011). The quiet and controversial: Ural family of *Mycobacterium tuberculosis*. Infect. Genet. Evol. *12*, 619-629.

S5. Pennarun, E., Kivisild, T., Metspalu, E., Metspalu, M., Reisberg, T., Moisan, J., Behar, D. M., Jones, S. C., and Villems, R. (2012). Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. BMC Evol. Biol. *12*, 234.

S6. Mokrousov, I., Vyazovaya, A., and Narvskaya, O. (2014). *Mycobacterium tuberculosis* Latin American-Mediterranean family and its sublineages in the light of robust evolutionary markers. J. Bacteriol. *196*, 1833–1841.