Mycobacterium tuberculosis: Unser afrikanischer Begleiter seit über 70‘000 Jahren

Tuberkulose - eine der weltweit tödlichsten Infektionskrankheiten - entstand vor über 70’000 Jahren in Afrika. Das zeigt eine neue genetische Untersuchung von 259 unterschiedlichen Tuberkulose-Bakterienstämmen aus der ganzen Welt. Der Studie zufolge, die am Sonntag 1. September in der Fachzeitschrift Nature Genetics erscheint, sind die TB Bakterien zusammen mit den ersten modernen Menschen aus Afrika ausgewandert und haben sich weltweit verbreitet. Die tödlichen Merkmale der Tuberkulose entwickelten sich in der engen Gemeinschaft mit dem Menschen und den sich verändernden Lebensumständen. Diese neuen evolutionären Erkenntnisse könnten die zukünftige Entwicklung neuer Medikamente und Impfstoffe massgeblich beeinflussen.

Multiresistente TB Mycobakterien (Mycobacterium Tuberculosis)

Tuberkulose (TB) ist weltweit noch immer eine der tödlichsten Infektionskrankheiten. Unbehandelt sterben 50% der an TB erkrankten Menschen. Und noch immer verenden daran jährlich ein bis zwei Millionen Menschen hauptsächlich in Entwicklungsländern. Die Entwicklung von Medikamentenresistenzen ist dabei eine ernsthafte Bedrohung im Kampf gegen die tödliche Krankheit.

Eine internationale Forschungszusammenarbeit unter Leitung von Sebastien Gagneux vom Schweizerischen Tropen- und Public Health Institut (Swiss TPH) hat den Zeitpunkt und den Entstehungsort der Tuberkulose bestimmen können. Mit Hilfe einer Vollgenomanalyse von 259 *Mycobacterium tuberculosis* Stämmen, die an verschiedenen Orten rund um den Globus gesammelt wurden, zeichneten die Forscher den genetischen Stammbaum der tödlichen Bakterien nach. Dieser Erbgutvergleich zeigt, dass die TB Bakterien vor mindestens 70’000 Jahren in Afrika entstanden sind.

Verblüffend enge Beziehung zwischen Mensch und *M. tuberculosis*
Die Forscher verglichen den evolutionären Stammbaum der TB Bakterien direkt mit jenem des modernen Menschen. Zur Überraschung der Forscher gleichen sich Form und die Verzweigungen der beiden Stammbäume örtlich und zeitlich in einem hohen Masse. „Die Evolutionswege des Menschen und der TB Bakterien zeigen auffällige Gemeinsamkeiten“, sagt Sebastien Gagneux. Daraus lässt sich auf eine sehr enge und direkte Beziehung der beiden Organismen schliessen.

Mensch und die TB Bakterien sind demnach nicht nur in derselben Weltgegend entstanden, sondern haben sich aus Afrika heraus gemeinschaftlich über die ganze Welt ausgebreitet. Die veränderte Lebensweise des modernen Menschen zunehmend in grösseren Gruppen und landwirtschaftlichen Strukturen während der so genannten Neolithischen Revolution hat günstigste Selektionsbedingungen für die direkte Mensch-zu-Mensch Übertragung der Bakterien geschaffen. Über die evolutionäre Auslese sind so immer tödlichere Keime entstanden. „Wir sehen, dass die Vielfalt der TB Bakterien sprunghaft ansteigt, just zum Zeitpunkt wo der Mensch aus Afrika ausgewandert ist.“, sagt Evolutionsbiologe Sebastien Gagneux.

Die neuen Resultate deuten auch darauf hin, dass TB im Unterschied zu anderen Infektionskrankheiten nicht von Haus- oder Nutztieren auf den Menschen übergesprungen ist. „TB Bakterien sind entstanden lange bevor der Mensch Tiere als Haus- und Nutztiere gehalten hat.“, sagt Sebastien Gagneux.

Neue Strategien für den Kampf gegen Tuberkulose
Tuberkulose bleibt eine globale Bedrohung. Neue Medikamente und Impfstoffe sind dringend notwendig um diese Infektionskrankheit zu bekämpfen. Mehrfach-Resistenzen gegen gängige TB-Medikamente steigen in Osteuropa, Asien und Teilen Afrikas dramatisch an. Die neuen grundlegenden Erkenntnisse zur Entstehungsgeschichte und Evolutionsbiologie der TB Bakterien könnten helfen, um neuartige Medikamente oder Impfstoffe und bessere Strategien zur Kontrolle der Krankheit entwickeln zu können.

Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans.
Iñaki Comas, Mireia Coscolla, Tao Luo, Sonia Borrell, Kathryn E Holt, Midori Kato-Maeda, Julian Parkhill, Bijaya Malla, Stefan Berg, Guy Thwaites, Dorothy Yeboah-Manu, Graham Bothamley, Jian Mei, Lanhai Wei, Stephen Bentley, Simon R Harris, Stefan Niemann, Roland Diel, Abraham Aseffa, Qian Gao, Douglas Young & Sebastien Gagneux. *Nature Genetics, AOP, Sept 1 2013. 10.1038/ng.2744*

Weitere Auskünfte Prof. Dr. Sebastien Gagneux

nature genetics

Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans

Iñaki Comas[1,2], Mireia Coscolla[3,4,23], Tao Luo[5,23], Sonia Borrell[3,4], Kathryn E Holt[6,7], Midori Kato-Maeda[8], Julian Parkhill[9], Bijaya Malla[3,4], Stefan Berg[10], Guy Thwaites[11,12], Dorothy Yeboah-Manu[13], Graham Bothamley[14], Jian Mei[15], Lanhai Wei[16], Stephen Bentley[9], Simon R Harris[9], Stefan Niemann[17], Roland Diel[18], Abraham Aseffa[19], Qian Gao[5], Douglas Young[20–22,24] & Sebastien Gagneux[3,4,24]

Tuberculosis caused 20% of all human deaths in the Western world between the seventeenth and nineteenth centuries and remains a cause of high mortality in developing countries. In analogy to other crowd diseases, the origin of human tuberculosis has been associated with the Neolithic Demographic Transition, but recent studies point to a much earlier origin. We analyzed the whole genomes of 259 *M. tuberculosis* complex (MTBC) strains and used this data set to characterize global diversity and to reconstruct the evolutionary history of this pathogen. Coalescent analyses indicate that MTBC emerged about 70,000 years ago, accompanied migrations of anatomically modern humans out of Africa and expanded as a consequence of increases in human population density during the Neolithic period. This long coevolutionary history is consistent with MTBC displaying characteristics indicative of adaptation to both low and high host densities.

Tuberculosis killed one in five adults in Europe and North America between the seventeenth and nineteenth centuries[1] and today remains a cause of high morbidity and mortality in much of the developing world[2]. Infectious diseases of humans can be divided into two broad categories[3]. Crowd diseases are generally highly virulent and depend on high host population densities to maximize pathogen transmission and reduce the risk of pathogen extinction through the exhaustion of susceptible hosts[4]. Many crowd diseases emerged during the Neolithic Demographic Transition (NDT) starting around 10,000 years ago, as the development of animal domestication increased the likelihood of zoonotic transfer of novel pathogens to humans and agricultural innovations supported increased population densities that helped sustain the infectious cycle[3]. In contrast, older human infections are often characterized by slow progression to disease, sometimes involving reactivation after many years of latent or asymptomatic infection; these characteristics have been proposed to reflect adaptation to low host population densities by allowing repletion of the reservoir of susceptible individuals[5]. Tuberculosis is reminiscent of a typical crowd disease in killing up to 50% of individuals when left untreated[3,6] and having evolved a mode of aerosol transmission that is promoted by high host densities. However, tuberculosis also displays a pattern of chronic progression, latency and reactivation that is characteristic of a pre-NDT disease[7]. Human tuberculosis was traditionally believed to have originated from animals[4], but more recent phylogenetic analyses of MTBC have suggested that the strains adapted to cause tuberculosis in animals diverged from the major human strains before NDT[8–13]. Moreover, human-associated MTBC is an obligate human pathogen with no known animal or environmental reservoir, suggesting that changes in human demography are likely to affect the evolution of MTBC. Here we used a population genomics approach to explore the evolutionary history of human MTBC, with a particular focus on the impact of changing host population sizes over time. Our results suggest a model that allows reconciliation of the apparent discrepancy between MTBC features characteristic of crowd diseases and those indicative of adaptation to low host densities.

[1]Genomics and Health Unit, Centre for Public Health Research (CSISP-FISABIO), Valencia, Spain. [2]CIBER (Centros de Investigación Biomédica en Red) in Epidemiology and Public Health, Barcelona, Spain. [3]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland. [4]University of Basel, Basel, Switzerland. [5]Key Laboratory of Medical Molecular Virology, Institutes of Biomedical Sciences and Institute of Medical Microbiology, Shanghai Medical College, Fudan University, Shanghai, China. [6]Department of Biochemistry and Molecular Biology, The University of Melbourne, Melbourne, Victoria, Australia. [7]Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Melbourne, Victoria, Australia. [8]Division of Pulmonary and Critical Care Medicine, University of California, San Francisco, San Francisco, California, USA. [9]Pathogen Genomics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [10]TB Research Group, Veterinary Laboratories Agency, Weybridge, New Haw and Addlestone, UK. [11]Department of Infectious Disease, King's College London, London, UK. [12]Centre for Clinical Infection and Diagnostics Research, King's College London, London, UK. [13]Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Ghana. [14]Department of Respiratory Medicine, Homerton University Hospital, London, UK. [15]Department of Tuberculosis Control, Shanghai Municipal Center for Disease Control and Prevention, Shanghai, China. [16]Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai, China. [17]Molecular Mycobacteriology, Research Center Borstel, Borstel, Germany. [18]Institute for Epidemiology, Schleswig-Holstein University Hospital, Kiel, Germany. [19]Armauer Hansen Research Institute, Addis Ababa, Ethiopia. [20]Medical Research Council (MRC) National Institute for Medical Research, Mill Hill, London, UK. [21]Department of Medicine, Imperial College London, London, UK. [22]Centre for Molecular Bacteriology and Infection, Imperial College London, London, UK. [23]These authors contributed equally to this work. [24]These authors jointly directed this work. Correspondence should be addressed to I.C. (inaki.comas@uv.es), Q.G. (qgao99@yahoo.com) or S.G. (sebastien.gagneux@unibas.ch).

Received 17 December 2012; accepted 1 August 2013; published online 1 September 2013; doi:10.1038/ng.2744

RESULTS

The global diversity of human-adapted MTBC

We sequenced the whole genomes of 186 strains representative of the global diversity of MTBC, combining these sequences with data from 34 already published strains and 39 additional newly sequenced strains corresponding to the lineage 2 'Beijing' family (**Supplementary Table 1**). In the global data set, after excluding repetitive and mobile elements, we identified 34,167 polymorphic sites (SNPs) (**Supplementary Table 2**), which we used to reconstruct phylogenetic relationships between these strains (**Fig. 1a**). This genome-based phylogeny was congruent with previous phylogenies based on other markers and resolved seven major lineages, with animal-adapted strains clustering together with the strains from lineage 6 (ref. 8). The phylogeny included the recently described lineage 7, which so far has only been observed in Ethiopia or in recent Ethiopian emigrants[14]. Principal-component analysis confirmed all main MTBC lineages and highlighted the close phylogenetic relationship between Eurasian lineages 2–4. These three lineages have collectively in the past been referred to as evolutionarily 'modern' (**Fig. 1b**) because of their comparably more derived position on the MTBC phylogeny and because they are thought to have spread more recently[8,11]. The maximum genetic distance between any 2 strains was 2,188 SNPs and involved a human and an animal strain and was 1,856 SNPs when only human clinical isolates were considered. Only 387 of the SNPs (1.1%) were homoplastic. Homoplasy can arise as a consequence of false-positive SNP calls because of positive selection or recurrent mutations, or because of recombination, as recently suggested[15]. However, the fact that only 1.1% of the sites were homoplastic supports the view that the population structure of MTBC is largely clonal, with little ongoing recombination occurring between strains[16,17].

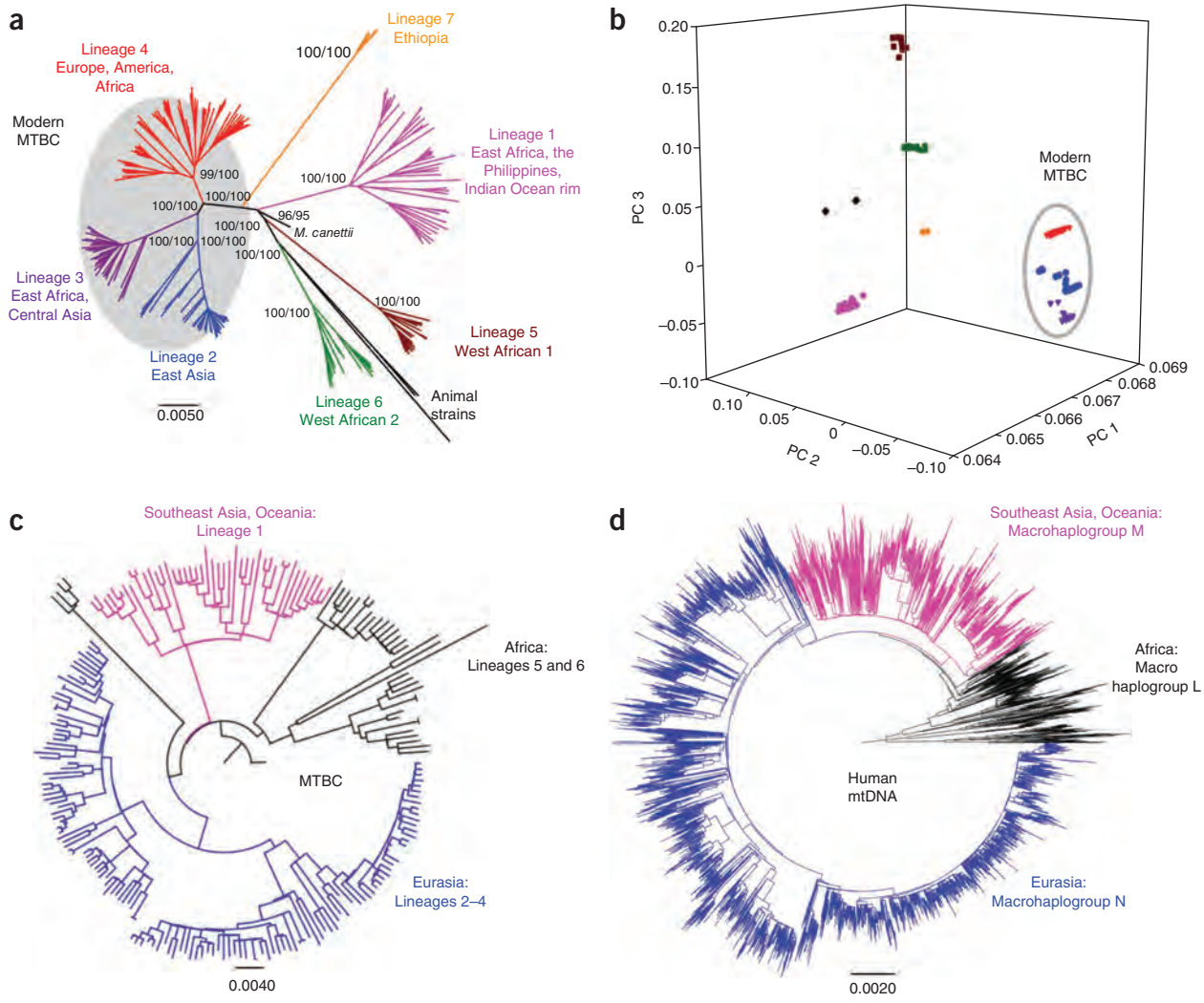

Figure 1 The genome-based phylogeny of MTBC mirrors that of human mitochondrial genomes. (**a**) Whole-genome phylogeny of 220 strains of MTBC. Support values for the main branches after inference with neighbor-joining (left) and maximum-likelihood (right) analyses are shown. (**b**) Principal-component analysis of the 34,167 SNPs. The first three principal-component axes (PC 1–PC 3) are shown; these discriminate between evolutionarily modern (gray circle) and ancient (all other) strains. Individual lineages are shown with the same colors as in **a**. (**c**,**d**) Comparison of the MTBC phylogeny (**c**) and a phylogeny derived from 4,955 mitochondrial genomes (mtDNA) representative of the main human haplogroups (**d**). Color coding highlights the similarities in tree topology and geographic distribution between MTBC strains and the main human mitochondrial macrohaplogroups (black, African clades: MTBC lineages 5 and 6, human mitochondrial macrohaplogroups L0–L3; pink, Southeast Asian and Oceanian clades: MTBC lineage 1, human mitochondrial macrohaplogroup M; blue, Eurasian clades: MTBC lineage 2–4, human mitochondrial macrohaplogroup N). MTBC lineage 7 has only been found in Ethiopia, and its correlation with any of the three main human haplogroups remains unclear. Scale bars indicate substitutions per site.

Table 1 Comparison of different dating scenarios for MTBC evolution

Dating scenario	MTBC-70	MTBC-185	MTBC-10	MTBC-65
Rationale	Emergence of MTBC with human mitochondrial DNA haplogroup L3	Emergence of MTBC with anatomically modern humans	Emergence of MTBC during NDT	Emergence of Out-of-Africa MTBC with human mitochondrial DNA haplogroup M
Dates inferred from models (in thousands of years ago)[a]				
MRCA of MTBC	73 (50–96)	198 (170–229)	11 (9–14)	67 (44–91)
Coalescent time for lineages 5 + 6	70 (48–88)[b]	184 (164–203)[b]	10 (8–12)[b]	61 (40–81)
Coalescent time for lineage 1	67 (46–88)	183 (160–207)	10 (8–12)	62 (42–82)[b]
Coalescent time for lineages 2–4	46 (31–61)	126 (104–148)	7 (6–10)	41 (26–55)
Period of maximum logistic growth	4–7	31–34	1	4–7
Substitution rate (SNPs per polymorphic site per thousand years)[c]	3.37×10^{-4} (2.38×10^{-4} to 4.65×10^{-4})	1.23×10^{-4} (1.04×10^{-4} to 1.46×10^{-4})	2.17×10^{-3} (1.71×10^{-3} to 2.68×10^{-3})	3.78×10^{-4} (2.62×10^{-4} to 5.36×10^{-4})

[a]Dates are shown as the median value and 95% HPD interval predicted in the corresponding Bayesian analysis. [b]Value provided as prior input in Bayesian analysis. [c]BEAST-predicted rate of SNP accumulation (per polymorphic position and thousand years). In the main text we use the estimated genomic substitution rate (per position per year) for comparative purposes with published estimations from other bacterial species.

African origin and codivergence of MTBC with modern humans

Several studies have proposed an African origin for MTBC[8,10,12]. We decided to formally test this hypothesis using our new whole-genome data. We used three independent phylogeographic analyses to determine the likely geographic origin of the most recent common ancestor (MRCA) of MTBC. Two different Bayesian analyses identified Africa as the most likely origin of MTBC, with East and West Africa showing combined posterior probabilities of 90% and 67%, respectively (**Supplementary Figs. 1–3**). Similarly, a maximum parsimony approach predicted 100% probability of an African origin. Taken together, these data support the hypothesis that MTBC originated in Africa.

Next, we sought to determine the putative age of the association between MTBC and its human host. Given that human-adapted MTBC is limited to humans and that both anatomically modern humans and MTBC originated in Africa, we tested whether MTBC and humans might have diverged in parallel; this would be particularly likely if the association between the two predates the NDT, as previously postulated[8,10,12]. To explore this possibility, we first compared our new MTBC phylogeny to a corresponding tree constructed from 4,955 mitochondrial genomes representative of the main human haplogroups (**Supplementary Table 3**)[18]. We observed striking similarities (**Fig. 1c**,**d**). In both cases, the early branching clades were found exclusively in Africa. Moreover, the trichotomy formed by the branching of the Out-of-Africa M and N mitochondrial macrohaplogroups from the L3 African source population was mirrored in the MTBC phylogeny by a similar relationship between lineage 1, Eurasian lineages 2–4 and African lineages 5 and 6. In addition to this qualitative similarity, comparison of the most common mitochondrial haplogroups with the most frequent MTBC lineages in the same country identified a strong quantitative association (by parsimony score and association index tests; $P < 0.01$ in all cases) (**Supplementary Fig. 4**, **Supplementary Table 4** and **Supplementary Note**). Taken together, these data are consistent with MTBC evolving in parallel with its human host.

Age of the association of MTBC and humans

Similarities in tree topology and phylogeographic distribution suggest that MTBC infected the early human populations of Africa. To further explore the association between MTBC and its human host, we tested for possible imprints of ancient human divergence times on the main phylogenetic lineages of MTBC using a Bayesian approach[19]. Several approaches have been used to date bacterial phylogenies (see refs. 20–22 for some examples). Unfortunately, none of these were applicable here because of the following reasons. First, although ancient DNA has been used to study the evolutionary history of other bacteria[20] and similar studies have been performed in tuberculosis in the past[23], no relevant whole-genome data are currently available for ancient DNA from MTBC strains. Second, although a mutation rate for MTBC has recently been estimated on the basis of a macaque infection model and molecular epidemiological data[24,25], it is well known that such short-term mutation rates cannot easily be extrapolated to the long-term substitution rates relevant for the time scale discussed here[26,27]. Third, and related to the previous point, although the isolation dates of some of the strains included in our analysis are known, at best they would allow the calculation only of a short-term mutation rate. Moreover, when performing a tip-to-date analysis of those strains ($N = 49$), we found that, in contrast to several other bacterial species[21,28–30], MTBC had no significant correlation between isolation time and phylogenetic divergence (correlation coefficient = 0.047).

Because of these limitations, we used an alternative approach to date our MTBC phylogeny. Specifically, we used as initial calibration points several key dates in human evolution. We tested three alternative models in which the coalescent time for the most basal MTBC lineages 5 and 6 was calibrated against (i) the emergence of anatomically modern humans 185,000 ± 20,000 years ago (MTBC-185)[31], (ii) the coalescent time of the L3 mitochondrial haplogroup 70,000 ± 10,000 years ago (MTBC-70)[32] and (iii) the beginning of the NDT 10,000 ± 2,000 years ago (MTBC-10)[3] (**Table 1**). We compared the timing of the branching points predicted by each of the models with estimated dates of known events in human history. A recent model based on the analysis of human whole-genome variation data sets suggests that the global dispersal of modern humans occurred through two major waves: an initial eastern dispersal around the Indian Ocean starting 62,000–75,000 years ago and a later dispersal into Eurasia 25,000–38,000 years ago[33]. Our MTBC-70 model showed a striking correlation with these human migration events by dating a first split of lineage 1 at 67,000 years ago (95% highest probability density (HPD) = 48,000–88,000 years ago), coinciding with the first wave of human migration[31], and a second split at 46,000 years ago (95% HPD = 31,000–61,000 years ago), matching the later dispersal throughout Eurasia (**Fig. 2a** and **Supplementary Fig. 5**)[34,35]. Coalescent dates for the branch leading to lineages 2 and 4 in the MTBC-70 model (30,000–46,000 years ago and 32,000–42,000 years ago, respectively) showed a good correlation with archaeological evidence of the presence of modern humans in Europe[35] and East Asia[36]. In contrast, our alternate model MTBC-185 postulated initial branching of Out-of-Africa lineages as early as 126,000–174,000 years ago when focusing

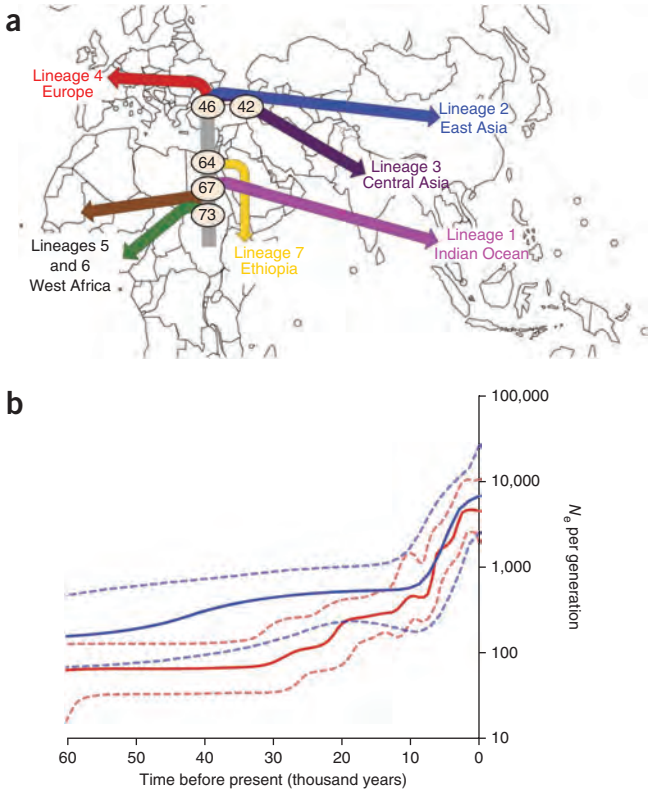

Figure 2 Out-of-Africa and Neolithic expansion of MTBC. (**a**) Map summarizing the results of the phylogeographic and dating analyses for MTBC. Color coding of lineages is the same as in **Figure 1a**. Major splits are annotated with the median value (in thousands of years) of the dating of the relevant node. Lineage 7 has so far been isolated exclusively in individuals with known country of origin in the Horn of Africa[14]. Lineage 7 diverged subsequent to the proposed Out-of-Africa migration of MTBC; it may have arisen among a human population that remained in Africa or a population that returned to Africa. (**b**) Bayesian skyline plots showing changes in population diversity of MTBC (red) and humans based on mitochondrial DNA (blue) over the last 60,000 years. Dashed lines represent the 95% HPD intervals for the estimated population sizes. N_e, effective population size.

on the branch leading to modern strains (**Supplementary Fig. 6**), which would suggest that the global dispersal of MTBC preceded that of anatomically modern humans. The MTBC-10 model, by definition, implies global dispersal within the last 10,000 years (**Supplementary Fig. 7**). Although MTBC has been spread by trade and conquest in recent centuries[8], the pattern of this dispersal does not match the phylogeographic distribution discussed above. Finally, a fourth model, MTBC-65, using the coalescent time of mitochondrial haplogroup M as a calibration time point for MTBC lineage 1, generated very similar results to the MTBC-70 model (**Table 1**). In summary, our phylogenetic analysis based on a 70,000-year time frame shows that MTBC has been infecting humans for at least the last 70,000 years.

Neolithic coexpansion of MTBC and humans

All the data presented so far strongly support the notion that human tuberculosis indeed predated NDT. How then could the features of tuberculosis typical of crowd diseases have arisen? To address this question, we used Bayesian skyline plots to estimate the changes in effective population size over time in the pathogen and human populations[19]. Analysis of our full MTBC sequence data set identified a main signal of population size increase starting about 10,000 years ago (**Fig. 2b**), suggesting that the expansion of MTBC occurred as a consequence of the increase in population densities that followed the establishment of the first human settlements during NDT[37] and not only because of a general increase in the total number of humans peopling the planet at the time. To test whether human population dynamics around that period coincided with those for MTBC, we used a data set previously described to maximize the information on human demographics during the Neolithic (**Supplementary Table 5**)[38]. The resulting skyline plot showed a Neolithic expansion of humans around 4,000–8,000 years ago (**Supplementary Fig. 8**), coinciding with the expansion of MTBC (Spearman's $R = 0.99$; $P < 0.00001$; **Fig. 2b** and **Supplementary Fig. 8**). Taken together, these findings indicate that the Neolithic period contributed to the success of MTBC, not by enhancing the likelihood of zoonotic transfer to humans as previously proposed, but because of combined increases in host population size and density.

The evolutionary history of MTBC on a regional scale

To analyze MTBC evolution at a regional level, we focused on lineage 2, which includes the Beijing family of strains. These strains have received particular attention because of their hypervirulence in laboratory models, their recent dissemination in human populations and their association with drug resistance[39]. Supplementing our global diversity set by sequencing the whole genomes of an additional 39 lineage 2 strains from China, we observed a strong correlation between skyline plots derived from lineage 2 genomes and a set of human mitochondrial genomes enriched for haplogroups from East Asia that likely originated just before, during or after the Neolithic

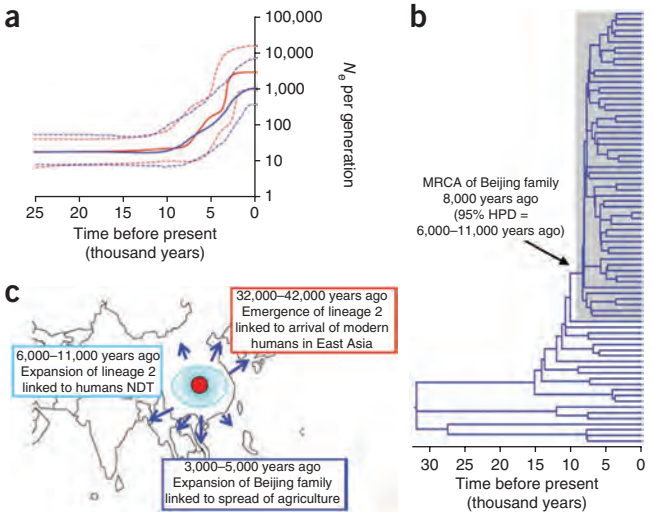

Figure 3 Neolithic expansion and spread of MTBC lineage 2 Beijing strains in East Asia. (**a**) Bayesian skyline plots indicating changes in lineage 2 diversity over time (red) compared with human mitochondrial DNA haplogroups from East Asia (blue). Dashed lines show 95% HPD intervals for the population size estimations. (**b**) Dated Bayesian phylogeny of MTBC lineage 2 based on coalescent analysis. (**c**) Map of the parallel origin and migration of MTBC and humans in East Asia, indicating the first archaeological evidence of modern humans in the region 32,000–42,000 years ago, coinciding with the migration of MTBC from central to East Asia, the start of the Neolithic period in the region indicated by the first evidence of domesticated crops in China coinciding with the origin of the MTBC Beijing family 8,000 years ago (HPD = 6,000–11,000 years ago) and the coexpansion of agriculture and the MTBC Beijing family into neighboring countries 3,000–5,000 years ago.

period (Spearman's $R = 0.97$; $P < 0.001$; **Fig. 3a** and **Supplementary Fig. 9**). MTBC-70 dating for lineage 2 is consistent with an initial arrival coincident with archaeological evidence of anatomically modern humans in East Asia[36] (32,000–42,000 years ago; **Supplementary Fig. 5**), a first expansion (6,000–11,000 years ago; **Fig. 3b**,**c**) alongside the emergence of agriculture in China 8,000 years ago[40] and a subsequent main expansion of the Beijing strains (3,000–5,000 years ago; **Supplementary Fig. 9**) coinciding with the spread of agriculture to neighboring regions (**Fig. 3b**,**c**)[37].

In summary, our data on the global and regional expansion of MTBC during NDT support the view that, although NDT was not the only period leading to large increases in human population sizes, it was the period where, in addition to human population growth, the densities of human populations increased following the first establishment of permanent human settlements. Hence, in addition to providing a springboard for global domination by modern humans, NDT was also central to the success of MTBC by generating growing numbers of susceptible hosts living under increasingly crowded conditions.

DISCUSSION

The common origin in Africa of MTBC and humans, the congruence in their phylogeographies and the dating of major branching events lead us to conclude that MTBC has been coevolving with anatomically modern humans for tens of thousands of years. The marked expansion of MTBC during NDT but not during earlier human expansion events[41,42] suggests that the success of this pathogen was primarily driven by increases in human host population density, which is typical of crowd diseases. However, the striking match between MTBC and human mitochondrial phylogenies supports a much older association between MTBC and its host and suggests that carriage of MTBC was ubiquitous in hunter-gatherer populations migrating out of Africa well before NDT. The fidelity of this match is unexpected. Considering the vulnerability and small numbers of human groups (some of today's hunter-gatherers live in groups of 20 or less[43]), it might have been anticipated that tuberculosis would have substantial detrimental impact on these groups and might therefore have precipitated its own extinction. In fact, the correspondence of the MTBC phylogeny with early human migration is strikingly similar to that observed with low-virulence *Helicobacter pylori*[44]. Perhaps latent infection with MTBC imparted some degree of immunity against more lethal pathogens encountered in the new environment or in contact with archaic human populations. Ongoing analyses of human microbiota highlight the fuzzy boundaries between commensalism and pathogenicity in health and disease[45]. A recent study has suggested that coinfection with *H. pylori* might protect against active tuberculosis disease[46]. Conversely, whether latent tuberculosis infection protects against gastric ulcers or stomach cancer caused by *H. pylori* in individuals infected with both bacteria is unknown but represents an intriguing possibility. In such a case, positive feedback between both infections would result in an asymptomatic individual benefiting from being infected by both bacterial species.

Alternatively, one could think of a model in which early populations carried the infection in a less virulent form, with transmission sustained by reactivation of disease in elderly individuals beyond the reproductive age. The possibility that disease characteristics might have changed over time as different MTBC populations were selected in different human societies may help to explain current epidemiological trends associated with increased dissemination of the Beijing family of MTBC[39] and decreased rates of disease caused by evolutionarily 'ancient' lineages of MTBC[47]. In addition to changes in population density, it can be anticipated that the pathology of tuberculosis during NDT would have been influenced by coinfections with novel crowd diseases and by variations in key nutrients such as vitamin D[48]. Similarly, it is important to consider the possibility of reciprocal adaptive changes to the human genome as a result of prolonged coevolution with MTBC[49].

In this study, we have compared MTBC phylogenic diversity to human diversity inferred from mitochondrial genome data. One advantage of using mitochondrial data is that these data have been used extensively to study recent human evolution. Furthermore, such data are available from almost any region of the world, and there is a large body of work studying human migrations that is based on the distribution of mitochondrial haplogroups. However, mitochondrial DNA is also limited in that it contains little phylogenetic information, and the existing data sets suffer from potential sampling bias. Increasingly, new DNA sequencing technologies are paving the way for studies of human diversity based on whole genomes[33]. Hence, in the context of a pathogen such as MTBC, future studies should be based on paired human and bacterial whole-genome information that is collected prospectively. Such an integrated approach will allow investigation of the molecular determinants of host-pathogen coevolution in human tuberculosis and other diseases.

The accumulation of more than 30,000 SNPs by human MTBC strains over the proposed time frame of 70,000 years corresponds to a long-term genome-wide substitution rate of 2.58×10^{-9} substitutions per site per year (95% HPD = 1.66×10^{-9} to 2.89×10^{-9}; **Table 1**). This rate is much lower than recent estimates of short-term substitution rates for experimental models and human outbreaks[24,25]. A decrease in substitution rates measured over increasing time intervals is a common feature of phylogenetic analyses[27], and an exponential decrease is observed in the substitution rate with time when we pool our data with those from other similar genome-based studies published recently (correlation coefficient = −0.9614; $P < 0.0001$; **Fig. 4**). Fixation or removal of single-nucleotide changes by natural selection can contribute to this phenomenon, although retention of a high proportion of nonsynonymous mutations suggests that

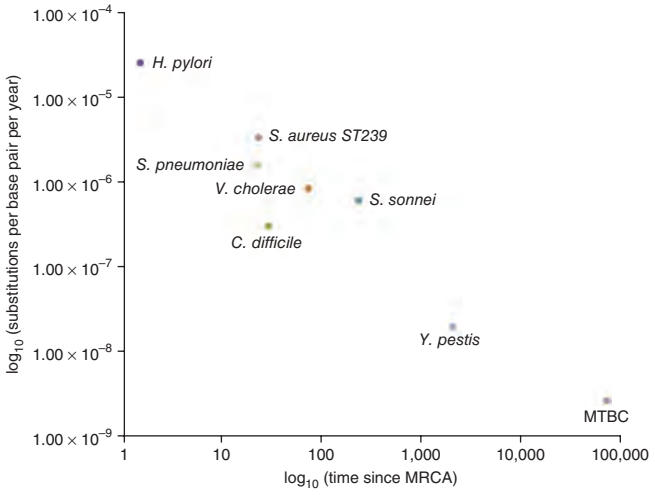

Figure 4 Time-dependent decay of substitution rates in bacteria based on whole-genome data sets. The scatter plot graph shows the relationship between substitution rate and time span between the MRCA and the last sampling date for each studied pathogen. Values were extracted from relevant publications that used whole-genome representative data sets and coalescent analysis of substitution rates (for a complete list of references, see the **Supplementary Note**).

natural selection has had a low impact on MTBC[8]. Alternative mechanisms to account for the reduction in genetic diversity over long time scales include serial founder effects linked to sequential expansions of human subpopulations and their associated pathogenic and commensal microbial flora[50].

In conclusion, we propose that MTBC has been a constant companion of anatomically modern humans during our evolution and global dissemination over the last 70,000 years. Furthermore, MTBC has been able to adapt to changing human populations. Exploration of changes that have occurred in this interaction over time may help predict future patterns of disease and to design rational strategies to bring an end to this historic partnership.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing reads for the previously unpublished genomes of strains that were used in this study (225 strains) have been deposited in the European Nucleotide Archive (ENA) under study number ERP001731. Additionally, we have analyzed the genomes from 34 previously published strains that are available at the Sequence Read Archive (SRA) under accessions SRS002426, SRS003212, SRS003328, SRS004666, SRS004753, SRS004754, SRS004756, SRS004757, SRS004758, SRS004759, SRS004760, SRS004761, SRS004762, SRS004763, SRS004764, SRS004831, SRS004841, SRS005175, SRS005448, SRS005450, SRS006765, SRS074557, SRS084142, SRX002001, SRX002002, SRX002003, SRX002004, SRX002005, SRX002429, ERS001592, ERS003236, ERS003237 and ERS003250). A complete list of the strains analyzed in this study together with sequencing and origin information is given in **Supplementary Table 1**.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Behar and S. Rosset for providing the mitochondrial genome sequences and C. Gignoux for advice on the mitochondrial Neolithic data set, N. Mistry (The Foundation for Medical Research) for providing bacterial strains and C. Dye, F. Balloux and L. Weinert for comments on the manuscript. This work was supported by the MRC UK (grants U.1175.02.002.00015.01 to S.G. and U117581288 to D.Y.), the Swiss National Science Foundation (PP0033-119205 to S.G.), the US National Institutes of Health (AI090928 and HHSN266200700022C to S.G.), the Leverhulme-Royal Society Africa Award (AA080019 to S.G.) and the Natural Science Foundation of China (grant 91231115 to Q.G.). DNA sequencing was partially supported by core funding of the Wellcome Trust (grant 098051) and by a Framework Programme 7 project of the European Community (SysteMTb HEALTH-F4-2010-241587 to D.Y.). I.C. is supported by European Union funding from the Marie Curie Framework Programme 7 actions (project 272086) and project BFU2011-24112 from the Ministerio de Economía y Competitividad (Spain).

AUTHOR CONTRIBUTIONS

I.C., Q.G., D.Y. and S.G. designed and supervised the study. M.C., S. Borrell, K.E.H., M.K.-M., J.P., B.M., S. Berg, G.T., D.Y.-M., G.B., J.M., L.W., S.R.H., S.N., R.D., A.A., Q.G. and S.G. provided MTBC strains and/or reagents. J.P., S. Bentley and S.R.H. contributed to the genome sequencing. I.C., M.C. and T.L. analyzed the data. I.C., M.C., T.L., S. Borrell, K.E.H., J.P., S. Berg, G.T., D.Y.-M., S. Bentley, S.R.H., S.N., A.A., Q.G., D.Y. and S.G. contributed to the manuscript writing. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Wilson, L.G. Commentary: medicine, population, and tuberculosis. *Int. J. Epidemiol.* **34**, 521–524 (2005).
2. World Health Organization. *The Global Plan to STOP TB 2011–2015* (World Health Organization, Geneva, 2011).
3. Wolfe, N.D., Dunavan, C.P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447**, 279–283 (2007).
4. Diamond, J. *Guns, Germs, and Steel: The Fates of Human Societies* 496 (W.W. Norton & Company, New York, 1999).
5. Blaser, M.J. & Kirschner, D. The equilibria that allow bacterial persistence in human hosts. *Nature* **449**, 843–849 (2007).
6. Berg, G. The prognosis of open pulmonary tuberculosis: a clinical-statistical analysis. *J. Am. Med. Assoc.* **114**, 1954–1955 (1940).
7. Barry, C.E. III *et al.* The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat. Rev. Microbiol.* **7**, 845–855 (2009).
8. Hershberg, R. *et al.* High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* **6**, e311 (2008).
9. Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A. & Behr, M.A. Genomic deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. *J. Infect. Dis.* **186**, 74–80 (2002).
10. Wirth, T. *et al.* Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog.* **4**, e1000160 (2008).
11. Brosch, R. *et al.* A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl. Acad. Sci. USA* **99**, 3684–3689 (2002).
12. Gutierrez, M.C. *et al.* Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Pathog.* **1**, e5 (2005).
13. Gagneux, S. *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **103**, 2869–2873 (2006).
14. Firdessa, R. *et al.* Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis in Ethiopia. *Emerg. Infect. Dis.* **19**, 460–463 (2013).
15. Namouchi, A., Didelot, X., Schöck, U., Gicquel, B. & Rocha, E.P.C. After the bottleneck: genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Res.* **22**, 721–734 (2012).
16. Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W. & Small, P.M. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc. Natl. Acad. Sci. USA* **101**, 4871–4876 (2004).
17. Comas, I. & Gagneux, S. The past and future of tuberculosis research. *PLoS Pathog.* **5**, e1000600 (2009).
18. Behar, D.M. *et al.* A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90**, 675–684 (2012).
19. Drummond, A.J., Ho, S.Y.W., Phillips, M.J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
20. Bos, K.I. *et al.* A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478**, 506–510 (2011).
21. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465 (2011).
22. Morelli, G. *et al.* *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat. Genet.* **42**, 1140–1143 (2010).
23. Djelouadji, Z., Raoult, D. & Drancourt, M. Palaeogenomics of *Mycobacterium tuberculosis*: epidemic bursts with a degrading genome. *Lancet Infect. Dis.* **11**, 641–650 (2011).
24. Ford, C.B. *et al.* Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat. Genet.* **43**, 482–486 (2011).
25. Walker, T.M. *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).
26. Morelli, G. *et al.* Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS Genet.* **6**, e1001036 (2010).
27. Ho, S.Y.W. *et al.* Time-dependent rates of molecular evolution. *Mol. Ecol.* **20**, 3087–3101 (2011).
28. Holt, K.E. *et al.* *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat. Genet.* **44**, 1056–1059 (2012).
29. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
30. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
31. Soares, P. *et al.* Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84**, 740–759 (2009).
32. Soares, P. *et al.* The expansion of mtDNA haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* **29**, 915–927 (2012).
33. Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94–98 (2011).
34. Henn, B.M., Cavalli-Sforza, L.L. & Feldman, M.W. The great human expansion. *Proc. Natl. Acad. Sci. USA* **109**, 17758–17764 (2012).
35. Stewart, J.R. & Stringer, C.B. Human evolution out of Africa: the role of refugia and climate change. *Science* **335**, 1317–1321 (2012).
36. Jin, L. & Su, B. Natives or immigrant: modern human origins in East Asia. *Nat. Rev. Genet.* **1**, 126–133 (2000).
37. Bellwood, P. & Oxenham, M. *The Neolithic Demographic Transition and its Consequences* 13–34 (Springer, New York, 2008).

38. Gignoux, C.R., Henn, B.M. & Mountain, J.L. Rapid, global demographic expansions after the origins of agriculture. *Proc. Natl. Acad. Sci. USA* **108**, 6044–6049 (2011).
39. Parwati, I., van Crevel, R. & van Soolingen, D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect. Dis.* **10**, 103–111 (2010).
40. Barton, L. *et al.* Agricultural origins and the isotopic identity of domestication in northern China. *Proc. Natl. Acad. Sci. USA* **106**, 5523–5528 (2009).
41. Atkinson, Q.D., Gray, R.D. & Drummond, A.J. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol. Biol. Evol.* **25**, 468–474 (2008).
42. Wei, W. *et al.* A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.* **23**, 388–395 (2013).
43. Hamilton, M.J., Milne, B.T., Walker, R.S., Burger, O. & Brown, J.H. The complex structure of hunter-gatherer social networks. *Proc. Biol. Sci.* **274**, 2195–2202 (2007).
44. Linz, B. *et al.* An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**, 915–918 (2007).
45. Littman, D.R. & Pamer, E.G. Role of the commensal microbiota in normal and pathogenic host immune responses. *Cell Host Microbe* **10**, 311–323 (2011).
46. Perry, S. *et al.* Infection with *Helicobacter pylori* is associated with protection against tuberculosis. *PLoS ONE* **5**, e8804 (2010).
47. de Jong, B.C. *et al.* Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *J. Infect. Dis.* **198**, 1037–1043 (2008).
48. Martineau, A.R. *et al.* Reciprocal seasonal variation in vitamin D status and tuberculosis notifications in Cape Town, South Africa. *Proc. Natl. Acad. Sci. USA* **108**, 19013–19017 (2011).
49. Barnes, I., Duda, A., Pybus, O.G. & Thomas, M.G. Ancient urbanization predicts genetic resistance to tuberculosis. *Evolution* **65**, 842–848 (2011).
50. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* **102**, 15942–15947 (2005).

ONLINE METHODS

Data sets. *MTBC data sets.* We have analyzed a total of 259 MTBC strains (including 1 *Mycobacterium canettii* strain used as the outgroup). We used two different strain sets for different aspects of the analyses.

(1) Global MTBC data set ($n = 220$). This data set represents a global collection of MTBC clinical strains covering all the known phylogenetic lineages of MTBC and including representatives from 46 countries. In addition, three strains from the animal-adapted lineage (including one strain of the *Mycobacterium bovis* BCG vaccine) were included as reference, and one strain of *M. canettii* was used as the outgroup. More detailed information can be found in **Supplementary Table 1**.
(2) MTBC lineage 2–enriched data set ($n = 75$). To explore the evolution of MTBC in a regional setting, we extended our collection of 36 MTBC strains from lineage 2 with an additional 39 strains that represent the population diversity of lineage 2 in China based on standard genotyping (**Supplementary Table 1**).

Illumina reads for the genomes of the new MTBC strains sequenced and described in this study have been deposited under project number ERP001731.

Human mitochondrial data set. For comparisons with human genetic diversity, we analyzed large data sets of complete mitochondrial genomes. There are limitations inherent to mitochondrial DNA. First, estimating the most frequent mitochondrial DNA haplogroup in a particular country is always difficult and is dependent on sampling. Second, mitochondrial DNA contains limited phylogenetic information. However, the reasons to focus on a mitochondrial marker rather than on a chromosomal marker include (i) the availability of information for most regions and countries in terms of mitochondrial DNA haplogroup frequencies and (ii) the possibility of comparison with previously published studies dealing with human mitochondrial DNA haplogroups, human migrations and population dynamics. We used three different sets of human mitochondrial genomes that were available in public repositories. These are listed in **Supplementary Tables 3**, **5** and **6**.

(1) Global reference data set of human mitochondrial DNA ($n = 4,955$). This data set is a compilation of most of the publicly available human mitochondrial genomes for which the haplogroup has been determined[18]. This data set includes representatives of most known human mitochondrial macrohaplogroups and derived haplogroups.
(2) Neolithic population expansion data set of human mitochondrial DNA ($n = 423$). This data set is derived from the data set reported by Gignoux *et al.*[38] and includes selected representative haplogroups known to have their origin either before, during or shortly after the Neolithic period. This data set is therefore maximized to detect signatures of population expansion around this period that could be obscured by earlier expansion events.
(3) East Asia–enriched Neolithic data set of human mitochondrial DNA ($n = 72$). For MTBC lineage 2, we complemented the data set for East Asia by adding any newly published human mitochondrial genome from the mitochondrial DNA haplogroups of interest (B4a1, F1a1, E1a and E1b).

Sequencing of MTBC strains. The majority of MTBC strains were sequenced during the present project at different sequencing centers (GATC (Germany), Wellcome Trust Sanger Institute (UK) and Southern Genome Center (China)); a few additional sequences were retrieved from publicly available databases. MTBC DNA was extracted using standard procedures. Single- or paired-end multiplexed Illumina sequencing was performed as described previously[51]. Briefly, sequencing was performed on a HiScanSQ instrument with TruSeq SBS kit HS chemistry (Illumina) to generate sequencing reads of between 51 and 100 bases in length, depending on the strain. Average genome coverage was 146.5 of the reference genome (strain-specific genome coverage is shown in **Supplementary Table 1**).

Mapping Illumina sequencing reads and SNP calling. Sequencing reads for each MTBC strain were mapped to the inferred MRCA of MTBC as previously determined[52] (the sequence of the MTBC MRCA is available upon request). We used two mapping approaches, the ungapped Mapping and Assembly with Quality (MAQ)[53] algorithm and the Burrows-Wheeler algorithm described in BWA[54], along with the MAQ SNP caller and SAMtools[55], respectively, to generate two different lists of SNPs. We kept those polymorphic positions called by both approaches (**Supplementary Table 7**). For a complete description of the SNP-calling procedure and annotation of the positions, see the **Supplementary Note**, as well as **Supplementary Figure 10** for a workflow of the SNP-calling procedure.

Phylogenetic and principal-component analyses. Human mitochondrial DNA data sets were obtained from the database of variant positions used by Behar *et al.*[18]. For the population expansion data set of human mitochondrial DNA during the Neolithic period, sequences from the relevant accessions described in Gignoux *et al.*[38] were downloaded, and genomes were aligned using the ClustalW[56] implementation in the BioEdit package[57] followed by manual curation. We removed the poorly aligned region known as the D-loop and kept polymorphic sites for subsequent phylogenetic and coalescent analyses. For the MTBC data sets, we used variable positions for all downstream analyses. In both cases, we applied phylogenetic distance as well as maximum-likelihood methods. For a complete description of the phylogenetic analyses, the identification of homoplastic sites and the principal-component analysis of the SNPs used, see the **Supplementary Note**.

Phylogeographic analyses. For the phylogeographic analyses, we used the BSSVS model implemented in BEAST 1.6 (ref. 58). We also used RASP[59], which implements both Bayesian and parsimony approaches to analyze the ancestral geographic ranges of MTBC lineages. We subdivided the world map into seven broad geographic areas and used them as a proxy for the most likely origin of each strain (see **Supplementary Fig. 1** for subdivisions and **Supplementary Table 1** for the origins of infected individuals). We used broad geographic areas instead of exact locations because the large number of locations to consider and, hence, the exchange rates to estimate would be unmanageable if using all individual countries. Predefined geographic areas were introduced for each MTBC strain according to the country of origin of the infected individuals. See the **Supplementary Note** for a complete description of the settings for the different phylogeographic analyses.

MTBC–mitochondrial DNA association test. We tested the hypothesis that modern lineages 2– 4, lineage 1 and the African lineages 5 and 6 are associated with the N, M and L human mitochondrial DNA lineages, respectively. To this end, we assigned for each MTBC strain from a given country a mitochondrial DNA haplogroup according to the frequency of the haplogroup in that country on the basis of a review of the published literature (**Supplementary Fig. 4**). Only the two most frequent MTBC lineages of a country and the two most frequent mitochondrial DNA haplogroups were considered, unless only one MTBC lineage occurred in the country, in which case it was assigned to the most frequent mitochondrial DNA of the country (**Supplementary Table 4**). We used BaTs (Bayesian Tip-association significance testing)[60] to test whether the main lineages of MTBC for each country tended to be associated with a particular human mitochondrial DNA macrohaplogroup (L, M or N) or haplogroup (A, B, D, E, F, G, H, K, L, M, R or U) (**Supplementary Fig. 4**, **Supplementary Table 4** and **Supplementary Note**). For the tests, we assumed that there was no MTBC lineage that corresponded with the L0, L1, L2 and L4 human mitochondrial lineages on the basis of the fact that no lineage 5 or 6 strains are found outside of West Africa where the human mitochondrial L3 haplogroup has the highest frequency[31]. However, even when we introduce L0, L1, L2 and L4, the test results did not change. BaTs implements two association indexes, a parsimony score that quantifies the number of state changes in the phylogeny (a low number indicates high clustering of states) and an association index that examines internal nodes and records the most frequent state in the taxa downstream of the node. A statistical test was carried out by reshuffling the various states across the phylogeny. Given the constrained phylogeographic distribution of lineages 5 and 6 (*Mycobacterium africanum*) to West Africa and their basal but close position to all the Out-of-Africa lineages, these *M. africanum* lineages correlate best with the human mitochondrial L3 haplogroup, which shows remarkable similarities.

doi:10.1038/ng.2744

BEAST analyses. We used BEAST v. 1.6 (ref. 19) to date the evolutionary events and population dynamics of MTBC and the human mitochondrial DNA haplogroups. BEAST implements the joint sampling of the posterior distribution of different evolutionary parameters, such as the substitution rate or the population size, under a coalescent framework. In all cases, we used a skyline plot before looking for changes in population size over time. For MTBC, we used two data sets. To explore different dating hypotheses, we used the complete MTBC data set, a total of 216 strains excluding the outgroup (*M. canettii*) and the animal-based strains. We used an uncorrelated log-normal distribution for the substitution rate in all cases. We imposed different prior values on the coalescent times of lineages 5 and 6 according to plausible time estimates. Because no fossil records or good substitution rate estimates are available for MTBC, we used this approach as a way to narrow down the origin and age of the extant strains of MTBC. We imposed normal distributions in the coalescent time of lineages 5 and 6, as time estimates for mitochondrial haplogroups are usually given in coalescent times and not in times of splitting events between groups: 185,000 (± 20,000), 70,000 (± 10,000) and 10,000 (± 2,000) years ago. We also added as a second anchor point the split of MTBC lineage 1 with a normal prior of 65,000 ± 10,000 years ago, based on the coincident geographic distribution of lineage 1 with human mitochondrial macrohaplogroup M. Similar approaches were followed to analyze mitochondrial DNA data sets, where we used both a molecular clock approach (by specifying a published substitution rate[31]) and a dating approach (by assuming that the height of the phylogeny was distributed normally around 185,000 years ago as a mean ± 20,000 years ago). Both approaches yielded similar results, and we report the results for the dating analyses. Similarly, for the East Asian clade, we specified priors for the age of the whole data set (60,000 ± 10,000 years ago) and for the individual haplogroups as described in the literature (B4a1, 11,000 ± 3,000 years ago; E1a, 9,000 ± 3,000 years ago; E1b, 6,000 ± 3,000 years ago)[18]. For a detailed description of the models and the statistical comparison of skyline plots, see the **Supplementary Note**.

51. Quail, M.A. *et al.* A large genome center 's improvements to the Illumina sequencing system. *Nature Methods* **5**, 1005–1010 (2008).
52. Comas, I. *et al.* Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat. Genet.* **42**, 498–503 (2010).
53. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
54. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
55. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
56. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
57. Hall, T.A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98 (1999).
58. Lemey, P., Rambaut, A., Drummond, A.J. & Suchard, M.A. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).
59. Yu, Y., Harris, A.J. & He, X. S-DIVA (Statistical Dispersal-Vicariance Analysis): a tool for inferring biogeographic histories. *Mol. Phylogenet. Evol.* **56**, 848–850 (2010).
60. Parker, J., Rambaut, A. & Pybus, O.G. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* **8**, 239–246 (2008).