The Evolutionary History, Demography, and Spread of the Mycobacterium tuberculosis Complex

MAXIME BARBIER and THIERRY WIRTH
Laboratoire Biologie Intégrative des Populations, Evolution Moléculaire; Institut de Systématique, Evolution, Biodiversité, UMR-CNRS 7205, Muséum National d’Histoire Naturelle, Univ. Pierre et Marie Curie, EPHE, Sorbonne Universités, 75231 Paris cedex 05, France

ABSTRACT With the advent of next-generation sequencing technology, the genotyping of clinical Mycobacterium tuberculosis strains went through a major breakup that dramatically improved the field of molecular epidemiology but also revolutionized our deep understanding of the M. tuberculosis complex evolutionary history. The intricate paths of the pathogen and its human host are reflected by a common geographical origin in Africa and strong biogeographical associations that largely reflect the past migration waves out of Africa. This long coevolutionary history is cardinal for our understanding of the host-pathogen dynamic, including past and ongoing demographic components, strains’ genetic background, as well as the immune system genetic architecture of the host. Coalescent- and Bayesian-based analyses allowed us to reconstruct population size changes of M. tuberculosis through time, to date the most recent common ancestor and the several phylogenetic lineages. This information will ultimately help us to understand the spread of the Beijing lineage, the rise of multidrug-resistant sublineages, or the fall of others in the light of socioeconomic events, antibiotic programs, or host population densities. If we leave the present and go through the looking glass, thanks to our ability to handle small degraded molecules combined with targeted capture, paleomicrobiology covering the Pleistocene era will possibly unravel lineage replacements, dig out extinct ones, and eventually ask for major revisions of the current model.

INTRODUCTION
Tuberculosis has plagued mankind over the centuries and probably accompanied modern Homo sapiens out of Africa. The epidemiological agent of phthisis, also known as “consumption,” reached its epidemic apex during the 18th and 19th centuries. During the industrialization era, the disease was associated with the concentration of labor and poor socioeconomic settings that ultimately favored the spread of this “crowd” pathogen. This high-burden period was then followed by a progressive decline of the death and disease tolls that predated the antibiotic era and the Mycobacterium bovis BCG vaccination. The evolutionary histories of the host and its pathogen are intricately associated, implying that tuberculosis can only be fully understood in the light of H. sapiens origins, migrations, and demography (1). Excluding these parameters from our analyses might lead us to false conclusions regarding evolution, epidemiology, and pathobiology. In the same line, there is also an urgent need to unravel the genomic features unknown to us.

Received: 13 January 2016, Accepted: 21 January 2016, Published: 12 August 2016
Editors: William R. Jacobs Jr., Howard Hughes Medical Institute, Albert Einstein School of Medicine, Bronx, NY 10461; Helen McShane, University of Oxford, Oxford OX3 7DQ, United Kingdom; Valerie Mizrahi, University of Cape Town, Rondebosch 7701, South Africa; Ian M. Orme, Colorado State University, Fort Collins, CO 80523
Correspondence: Maxime Barbier, maxime.barbier@etu.ephe.fr © 2016 American Society for Microbiology. All rights reserved.
that can explain the contrasted infectivity and transmission observed between *Mycobacterium tuberculosis* complex (MTBC) lineages (2–4), without neglecting the genetic architecture of the host’s immune system (3).

Another challenge we have to face is the effect of globalization, i.e., the dramatic increase of population and individual movements that encompass touristic activities, refugee diasporas, and, soon to come, climatic migrants. This ongoing maelstrom has multiple consequences, such as an increasing number of patients infected by nonendemic strains, the spread of multidrug-resistant (MDR) strains from health-care-deficient countries, and the frightening specter of the expansion of totally drug-resistant (TDR) strains (6). In this review, we will illustrate how genomic insights driven by whole-genome sequencing and comparative genomics can help us to combat this old foe, and unravel its evolutionary history, spread, and demography. From a more practical point of view, the approaches we will discuss here, combined with selection and population genomics models, might also help us to evaluate the impacts of treatment programs on the relative transmission success, to pinpoint the molecular targets of selection, and, eventually, to develop new drugs.

**HISTORICAL CONSIDERATION AND EARLY (MIS)CONCEPTIONS ON TUBERCULOSIS EVOLUTION**

Few diseases, with the exception of plague (*Yersinia pestis*), have left such an important written signature as tuberculosis. The first literary traces were detected in Chinese medical texts predating the Xia dynasty and in the Indian Vedas (7), respectively, some 5,700 and 3,500 years ago. Until recently, little was known about tuberculosis origins, evolution, and spread. Thanks to the development of molecular tools, four distinct species were identified as causing the disease: *M. tuberculosis*, the human pathogen; *M. bovis* (8), found primarily in cattle; *Mycobacterium africanum* (9), isolated from African patients; and *Mycobacterium microti* (10), isolated from voles. These species were defined based on the host from which strains had been isolated. However, biochemical analyses, including *in vitro* growth rates, microscopic observations, and differential host-specific pathogenicity, suggested that interspecies borders were less well defined than initially expected (11). All these taxa belong to the MTBC, although their status in terms of taxonomic level (species, subspecies) might be further debated. In this group, *M. tuberculosis* and *M. bovis* are the more prevalent ones, although this might be due to strong sampling biases driven by health-economic priorities, as well as by differences in access to funding. *M. tuberculosis sensu stricto* infects humans and, until recently, the species was divided into five variants based on biochemical properties, namely the classical human, Asian human, bovine, African I, and African II variants (12).

The initial paradigm concerning the evolution of *M. tuberculosis* was that the bacillus evolved from *M. bovis* (13). Thanks to novel molecular data, however, this scenario was revised (14), although the old concept keeps being cited (15). The observations that led to these prime conclusions were the following. First, many diseases afflicting humans are zoonoses, and tuberculosis should be no exception. Famous examples of transmission from animals to humans encompass the Ebola virus, HIV, and Chagas’ disease (13). The transmission process can oscillate between sporadic outbreaks with little human-to-human transmission and a more settled coevolution if the bug can adapt to its new niche. Based on this knowledge, the initial hypothesis was that a cattle *M. bovis* strain infected a human and successfully spread in the *H. sapiens* populations. After some millennia of coevolution, the bacterium specialized to its novel host, became human specific, and is now known as *M. tuberculosis*. In fact, it is not unusual to see patients infected by bovine tuberculosis, with transmission occurring via aerosols or the consumption of infected milk. Moreover, until now, no human remains older than 11,000 years have shown traces of tuberculosis disease (16, 17), whereas the most ancient animal case has been found in a 17,000-year-old extinct bison (18).

This apparent anteriority of animal infection was used to promote the cattle-to-human transmission route hypothesis. Furthermore, the fact that the earliest human remains carrying tuberculosis date back to the Neolithic revolution (8,000 to 10,000 years ago) is intriguing and suggests some causality with the rise of domestication. It is tempting to think that the concomitant increase of animal stocks and host population size favored the interactions and contacts between these two players. Indeed, in the past, humans shared their home with bovines to protect them against predators and extreme temperatures (14): a single infected and coughing animal might have been able to transmit the disease to an entire family.

Here we see the dangerous attraction we have for nice, logical, flowing narratives; yet, mycobacterial interspersed repetitive unit genotyping and whole-genome sequencing (WGS) provided strong evidence against such a linear explanation, as we shall see in the next section.
THE PREGENOMIC ERA AND FIRST-GENERATION PHYLOGENETIC ANALYSES

The Fingerprint Era

The advances of molecular biology enabled the study of bacterial DNA, unraveled fine-scale genetic structures, and clearly segregated sister strains, which previously seemed nearly identical, mostly because morphological and biochemical traits provide little information about relatedness and species phylogenies. We will present first the main pre-next-generation sequencing (NGS) methods that enabled us to discriminate the principal MTBC families and to disentangle their evolutionary link, and how these methods shifted our vision of tuberculosis evolution and spread.

One of the first typing techniques applied to M. tuberculosis was the restriction fragment length polymorphism (RFLP) method. It is a fingerprint-based approach that relies on the enzymatic digestion of the circular chromosomal DNA, followed by gel electrophoresis revealed with radiolabeled probes targeting a particular sequence, such as an insertion sequence (IS). The IS6100 RFLP analysis (19) has been widely used for M. tuberculosis molecular typing. The IS6100 sequence is usually present in multiple copies on the chromosome. Depending on the locations and the copy number of this element, a profile is established that allows for strain discrimination. Such profiles enabled the demonstration that M. microti was responsible for human infections (20). Other insertion sequences have been used such as IS1081 (21) and IS986 (22), but they did not reach the success of IS6100. This marker presents a high evolutionary pace and therefore evolves relatively quickly in an otherwise relatively homogeneous genetic background. Therefore, IS6100 proved very useful in epidemiological studies and facilitated the segregation of clusters of closely related strains or, in the best case, of clones (23). A major drawback of this technique comes from its poor portability and laboratory dependency in terms of fingerprint profiles, leading therefore to little insight at larger evolutionary scales (24). Besides, differentiation of strains is strongly dependent on the number of IS6100 copies. Strains with high copy numbers are accurately differentiated from their close variants, while strains with few copies are more difficult to segregate. Numerous other markers have been used in the field of microbiology with more or less the same advantages and flaws (25). Another limitation of RFLP is that it requires mycobacterial culture, lasting from 20 to 40 days. This time frame is very long when studying infection chains in a clinical context. According to a search on the Web of Science database, the topic “IS6100” reached a citation apex in 2012 and now follows a gentle but regular decline.

The Multilocus Era

Next came PCR-based techniques that allowed fast, reproducible, and efficient typing (between 1 day and 1 week) like the spacer oligotyping method, called “spoligotyping” (26). The goal of this technique is to type the direct repeat (DR) locus of M. tuberculosis. This locus is an alternation of DRs, composed of a well-conserved sequence of 36 bp, and nonrepetitive spacer sequences, 34 to 41 bp long. M. tuberculosis strains can be discriminated based on their number of DRs and the presence or absence of particular spacers (27). Spoligotyping is therefore an efficient typing method (28) that differentiates MTBC strains from other environmental mycobacteria and clearly separates M. bovis from M. tuberculosis. It has less discriminatory power than IS6100, when present in high copy numbers, but it is present in all strains, unlike IS6100. The principal inconvenience of this method is that spoligo patterns are CRISPR structures that play a role in Eubacteria and Archaebacteria defense against phages (29). Consequently, this marker is under strong diversifying selection, prone to homoplasy, and of little interest for phylogenetic reconstructions, if any.

The second PCR-based method developed in the early 2000s is a high-resolution typing method based on variable number tandem repeats (VNTRs) of genetic elements named mycobacterial interspersed repetitive units (MIRUs) (30). Those markers resemble the human minisatellite-like regions developed by Sir Alec Jeffreys and widely used in forensics. MIRU loci are scattered in the genome of M. tuberculosis and consist of repetitive patterns 51 to 77 bp long. Only minor indels or polymorphisms occur in these sequences, mostly following the so-called stepwise-mutation model (SMM), meaning that the allelic state changes by the acquisition or the loss of one repetitive unit. The typing of these MIRUs is simply the measure of the number of repetitions at each locus where the number of repetitions varies between 0 and a maximum of 25. The MIRU typing underwent progressive upgrading steps, from 12 to 15 and finally 24 MIRU loci.

A third way used to differentiate strains consisted of the sequencing of a set of structural genes, allowing defining their relatedness based on sequence polymorphisms. Sreevatsan et al. implemented this approach in 1997 (31). They used 26 different genes or gene fragments, in which they observed a lack of neutral mutation, with up to 95% of nonsynonymous mutations associated with antibiotic
resistance. However, phylogenetic inferences and evolutionary scenarios inferred from genes under strong positive selection (involved in antibiotic resistance, for example) are generally not reliable. Alternatively, taking into account the sole synonymous single nucleotide polymorphisms (SNPs) might be a solution but this remains very restrictive. The authors ultimately used the only two nonsynonymous mutations that were not involved in antibiotic resistance to define three different groups. We see here that, because of technical barriers, biased locus choice, and little genetic diversity of the MTBC, the pre-NGS area sequencing studies remained tricky and led to partially misleading conclusions.

Yet another simple analytical approach turned out to be far more promising. Thanks to genome-wide comparisons (32), variable regions resulting from insertion-deletion events have been discovered. Approximately 20 such regions are phylogenetically highly informative since they follow a completely parsimonious and non-homoplastic evolutionary path, from presence to definitive loss, turning any change into a marble-engraved event (33). Based on these so-called regions of difference (RDs), a new evolutionary scenario emerged that contradicted previous thoughts (Fig. 1) because it stated that human strains did not derive from M. bovis. RD-based analyses even suggested that humans transmitted

**FIGURE 1.** Diagram of the proposed evolutionary pathway of the tubercle bacilli illustrating successive losses of DNA in certain lineages (gray boxes). The diagram is based on the presence or absence of conserved deleted regions and on sequence polymorphisms in five selected genes. The distances between certain branches may not correspond to actual phylogenetic differences calculated by other methods. Blue arrows indicate that strains are characterized by katG463. CTG (Leu), gyrA95 ACC (Thr), typical for group 1 organisms. Green arrows indicate that strains belong to group 2 characterized by katG463 CGG (Arg), gyrA95 ACC (Thr). The red arrow indicates that strains belong to group 3, characterized by katG463 CGG (Arg), gyrA95 AGC (Ser), as defined by Sreevatsan et al. (31). Adapted from Brosch et al. (33).
tuberculosis to cattle and other animals rather than the other way around. One deletion, TbD1, separates “modern strains” from “ancient strains.” This latter clade comprises animal strains, *M. africanum*, and some less virulent human strains, whereas “modern strains” exclusively infect humans. An interesting analysis of PhoPR virulence factors provided some novel insights into those splits and might explain how some mutations lowered the virulence of strains belonging to *M. bovis* and *M. africanum* (34). As an alternative to overstep the lack of accuracy or the drawbacks of all markers previously presented, some researchers chose to combine them (35). Such combined analyses have been conducted in numerous surveys and allowed the separation of *M. tuberculosis* human strains into different clades (Table 1). Those clades were initially named based on the prevalence and geographical source of their members (36, 37). The most remarkable phylogeographical clades belonging to the “modern strains” were Beijing (highly prevalent in East Asia), CAS (central Asia), and Haarlem, X, and LAM (Latin American-Mediterranean). *M. africanum* and EAI (East African Indian) composed the “ancient strains” group. Captivatingly, lineages from neighboring regions are more closely related than randomly chosen lineages, advocating for a strong biogeographical structuring: Haarlem, X, and LAM clades are more prevalent in Europe and cluster together; the same holds for the Asiatic Beijing and CAS clades. The other lineages form a paraphyletic group of ancient strains, which are essentially restricted to Africa and India. The observed relationships between MTBC lineages are similar to those observed in humans, suggesting that humans could have carried tuberculosis for millennia and that the present-day geographical distribution of tuberculosis has been shaped by ancient if not first human migrations out of Africa. The hypothesis that humans and *M. tuberculosis* coevolved and spread together has been studied and detailed by Wirth et al. (38). Using MIRU genetic markers, the authors identified two major clades, one composed of human strains only and one containing both human and animal strains. Interestingly, the basal and genetically more diverse lineage of the second clade infects humans, confirming that animal strains derived from human ones. Moreover, using Bayesian approaches and coalescent-based theory, they estimated the clade ages and inferred the *M. tuberculosis* demographic history. Based on these calculations, the common ancestor of the MTBC appeared some 40,000 years ago. In a second step, the ancestral strains reached the Fertile Crescent where they diversified during and shortly after the onset of domestication, 10,000 years ago. They ultimately spread out of Mesopotamia, accompanying different human migration waves in Africa, Asia, and Europe, and gave rise to locally adapted pathogens. Furthermore, a strong signal of demographic expansion was detected in the past 200 years, concomitant with industrialization. All these clues point toward a strong association and long coevolution between *H. sapiens* and *M. tuberculosis*.

Last, just before the rise of NGS and WGS, some researchers began to use SNP-based approaches to assess lineage relationships and to unravel deep MTBC sublineages. Since mutations are rare in *M. tuberculosis* genomes, they compared the complete genomes of few available reference strains and identified a list of SNPs. Then they sequenced these genes or called the SNPs in a large NGS data set gathered from strain collections (39–41). The authors retrieved the principal clades described above, but all generated phylogenies turned out to be poorly resolved, ending in star-like topologies. At first glance, one might have invoked a sudden radiative burst and a hard polytomy. What we were facing was, in fact, a methodological issue called ascertainment bias, which is often driven by biased taxonomic sampling (42). Indeed, strong ascertainment bias and related phylogenetic reconstructions systematically lead to the collapse of divergent lineages into single points, failing therefore to generate reliable tree topologies. This is exemplified by the Filliol et al. (41) paper, where the authors reached the unrealistic conclusions that *M. tuberculosis* had an Indian origin.

### NGS AND TUBERCULOSIS

#### EVOLUTIONARY HISTORY

**The Global Picture**

The ultimate knowledge that can be gathered using NGS is a complete list of all nucleotides that constitute the circular chromosome of a strain. Our understanding of the evolutionary relationships of the different MTB lineages, their radiation, and time to the most common ancestor (TMRCA) greatly profited from WGS, resulting in a quantum-leap progress in the field of MTBC phylogenetics (43). NGS favored the characterization of the genetic diversity of an increasing number of strains, covering different lineages and large geographic distributions. Thanks to a high-quality reference genome (44) (Sanger sequenced) gathered from the H37Rv laboratory strain, the scientific community has a template on which Illumina or Roche 454 reads can be mapped. Unraveling the topology of the MTBC tree is also highly dependent on the availability of a reliable outgroup.
## TABLE 1 Correspondence table of the MTBC human-adapted strains identified by main typing methods and including the latest nomenclature

<table>
<thead>
<tr>
<th>Evolutionary age (species)</th>
<th>Lineage name based on LSP/SNP&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Lineage and sublineage [RD associated]</th>
<th>Spoligotype family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancient lineage (M. tuberculosis)</td>
<td>Indo-Oceanic lineage 1 [RD239]</td>
<td>1.1 1.1.1 1.1.2 1.1.3 1.2 1.2.1 1.2.2</td>
<td>EAI4 and EAI5</td>
</tr>
<tr>
<td>Modern lineages (M. tuberculosis)</td>
<td>East-Asian lineage 2 2.1 (non-Beijing)</td>
<td>2.1.1 [RD150] 2.1.2 [RD142]</td>
<td>Beijing</td>
</tr>
<tr>
<td></td>
<td>East African-Indian lineage 3 [RD750]</td>
<td>3.1 3.1.1 3.1.2 3.1.2.1 3.1.2.2</td>
<td>CAS except CAS1-Dehi CAS1-Kili</td>
</tr>
<tr>
<td></td>
<td>Euro-American lineage 4 4.1</td>
<td>4.1.1 (X-type) 4.1.2</td>
<td>X2 X3 and X1 T1 and H1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.2 4.2.1 (Ural) 4.2.2</td>
<td>H3 and H4 LAM7-TUR and T1 LAM7-TUR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.3 (LAM) 4.3.1 4.3.2</td>
<td>LAM9 LAM3 LAM3 LAM9 and T5 LAM1 LAM11-ZWE, LAM9, LAM1, and LAM4 LAM11-ZWE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.4 4.4.1 4.4.1.1 (S-type) 4.4.1.2</td>
<td>S T1 T1 and T2 H3, H4, and T1 T2-Uganda</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.5 [RD122] 4.6</td>
<td>T1 and T2 T2-Uganda</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T1 and T5 T1, T2, T3, T4 and T5</td>
</tr>
<tr>
<td></td>
<td>4.7 4.8 [RD219]</td>
<td>4.7.1 4.7.2</td>
<td>AFRI&lt;sub&gt;2&lt;/sub&gt; and AFRI&lt;sub&gt;3&lt;/sub&gt; AFRI&lt;sub&gt;1&lt;/sub&gt;</td>
</tr>
<tr>
<td>Ancient lineages (M. africanum)</td>
<td>West-Africa lineage 1 5 [RD711]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediary lineage (M. tuberculosis)</td>
<td>West-Africa lineage 2 6 [RD702]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lineage 7 7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Regions of deletion (RD) are given in brackets and appear below the lineage/sublineage in which they are present. Synthetic table adapted from Coll et al. (86).

<sup>b</sup>LSP, large sequence polymorphism.
Fortunately, Supply and colleagues (45) sequenced and analyzed the whole genomes of five strains belonging to the smooth tubercle bacilli (STB), the closest outgroup known so far (46, 47). These strains harbor a unique smooth colony phenotype on culture media, are less persistent and virulent than their M. tuberculosis counterpart, and were essentially collected from the Horn of Africa, the cradle of humankind. Furthermore, the so-called “Mycobacterium canettii” and/or “Mycobacterium prototuberculosis” strains display a unique feature in the MTBC world: they are highly recombinogenic and they are prone to horizontal gene transfer (HGT). Indeed, they possess distinct CRISPR-Cas systems relative to M. tuberculosis that are closely related to the genera Thioalkalivibrio, Moorella, and Thiorhodovibrio (45). Interestingly, this latter species is adapted to warm and salty waters, a type of environment that is often encountered in the western part of Djibouti where large saline lakes coexist with hot springs. These scars of past genetic exchanges definitively advocate for an environmental origin of the ancestor of M. tuberculosis that might date back 3 million years (46).

Once rooted with M. canettii, the first attempt to solve the evolutionary history of the MTBC with full genomes relied on a set of 25 M. tuberculosis strains representing the six main human lineages known at that time (48). The molecular diversity of those strains remained rather modest, with only one SNP call for every 3 kb of sequence generated, highlighting the relative youth of this human pathogen, its clonality, and putative rise through a major bottleneck. The neighboring tree proposed by the authors did not add much in terms of branching order but illustrated the power of genomics in terms of bootstrap branch supports (≥99%) and within-lineage resolution. Three major clades could be distinguished, one encompassing lineages 2, 3, and 4; followed by its sister group, lineage 1; and, finally, a marginally more basal group represented by the two M. africanum lineages (Table 1). Because the rationale of this first genomic paper was to study M. tuberculosis human T-cell epitopes, the absence of animal strains was not surprising. However, in terms of evolutionary history, this no-attendance needed to be corrected in future studies. This was done in a landmark study (49) where genomes of 239 M. tuberculosis strains were analyzed. At such scales, with more than 30 strains per major lineage, the likelihood to get much closer to the real picture significantly increases. Comas and colleagues included in this study a new member of the MTBC, the recently described lineage 7 (50), which was only collected from Ethiopian patients, as well as a couple of animal strains. Again, the maximum-likelihood tree confirmed the monophyly of the modern strains (L2, 3, and 4), but also suggested that the animal lineage diverged from the African lineage 6 (M. africanum). Another important feature is the strong biogeographical structure of the different lineages; their distribution around the planet is not random at all, and clearly corresponds to well-defined geographic and cultural areas. This observation, coupled with the fact that tree topologies and geographic distribution between MTBC strains and the main human mitochondrial macrohaplogroups were highly similar, prompted the authors to calibrate the tuberculosis evolutionary tree on its human backbone. More specifically, the striking resemblance and branching order of the Southeast Asian and Oceania tuberculosis strains and the Southeast Asian, Oceanian macrohaplogroup M in humans were used for this purpose (Fig. 2).

This elegant approach, coupled with a coalescent-based approach, indicates that the MTBC emerged at least 70,000 years ago. The demographic success and the timing of the propagation of M. tuberculosis were evaluated with Bayesian skyline plots (51–53) that unraveled the effective population size of the bug through time. According to this scenario, MTBC accompanied the migrations of anatomically modern humans out of Africa and started to spread at a higher pace during the Neolithic demographic transition (54). It is tempting to connect a sustainable infectious cycle with the advance of farming and domestication, accompanied by dramatic changes in lifestyle, from hunter-gatherers to farmers, from low-density populations to local crowds. However, the data also show that the conquest of the Indian Ocean areas by lineage 1 largely predated the Fertile Crescent onset, starting as early as 67,000 years ago and followed by a second wave of peopling that reached the Middle East, Europe, and Asia some 46,000 years ago. Overall, WGS highlights the coevolution between a host and its pathogen, H. sapiens and M. tuberculosis, their intricate evolutionary histories, their African origin, and their adaptation from low to high population densities.

Yet recently a new publication dramatically affected the temporal dimension of the scenario presented above. Bos et al. (55) analyzed three 1,000-year-old mycobacterial genomes from Peruvian skeletons showing stigmata of tuberculosis infection that proved to be Mycobacterium pinnipedii (Fig. 3), a type of strain mostly isolated from seal species in the Southern Hemisphere. It is worth mentioning that two of the archeological sites (El Algodonal and Chiribaya Alta) were close to the Rio
The team led by German experts in ancient DNA managed to successfully sequence these genomes by applying DNA capture (56) and genomic assembly of the metagenomic reads. The assembled genomes harbored the typical signature and damage of ancient DNA, and accounted for 2% of the total reads. The authors then calibrated the molecular clock using the archeological data and the fact that branch lengths are a function of the elapsed time, being longer for strains collected in the 21st century and shorter for much older strains. This Bayesian calibration process, under a relaxed clock model, resulted in a substitution rate of about $5 \times 10^{-8}$ substitutions per site per year, placing the most recent common ancestor for the MTBC at 4,000 years, which turns to be more than one order of magnitude younger than the age proposed by Comas and colleagues (49).

For comparative purposes, we should mention that the substitution rate obtained by Comas et al. was much slower, i.e., $2.6 \times 10^{-9}$ substitutions per site per year. This MTBC TMRCA dating issue definitively splits the mycobacteriology community into two entities, i.e., the pros and the cons. The more recent dating of the Bos study conflicts with numerous archeological proofs, including evidence of MTBC in a 17,500-year-old bison in Wyoming, United States (18), the presence of a 9,000-year-old modern tuberculosis strain in a Neolithic infant skeleton from Israel (16), and an animal MTBC strain harboring the RD9 deletion some 7,000 years ago (57).

The cumulative evidence gathered from amplification of IS6100 and spoligotyping patterns from bones predating the Bos et al. MTBC TMRCA are questioned by some scientists, claiming that these mobile genetic elements and CRISPR systems are not MTBC specific enough and that they might be observed in environmental mycobacteria (58, 59), leading to false positives. The same arguments are used to question the validity of the presence of mycolic acids (60, 61) in biosamples to identify MTBC strains. Other colleagues came to the same molecular clock as Bos et al. (55) using a calibration point based on aboriginal communities in Canada that acquired $M. \text{tuberculosis}$ via the fur trade in the late 18th century (62). Nearly identical rates were obtained again based on a calibration relying on an 18th-century mummy collected from a Dominican church in Hungary (63). Furthermore, Pepperell et al. (62) did not find statistical support for codivergence of $M. \text{tuberculosis}$ with its host in formal phylogenetic congruence tests.
Another by-product of this study is that, according to the authors, seals are the source of New World human tuberculosis, therefore predating the likely entry of tuberculosis in South America with the Conquistadores, putatively harboring lineage 4 strains in their lungs. The Bos et al. conclusions based on the sole observation of a couple of ancient Peruvian humans are overinterpreted in the best case if not dubious at all. This scenario possibly transforms the exception into the rule by extrapolating conclusions based on local observations to continental-scale lessons. A more parsimonious setup could be built on a small population of indigenous hunters who incidentally contracted tuberculosis from infected seals that might be part of their natural prey or diet. Such transfers are rare but can be observed in zoos; notably, South American sea lions managed to infect a camel and a Malayan tapir in neighboring enclosures with their \textit{M. pinnipedii} strains (64), and animal keepers were infected in a zoo in the Netherlands (65). The observation made at the southern border of Peru might be therefore anecdotal and may have resulted in an evolutionary dead end. Before claiming that \textit{M. pinnipedii} plagued South and North American indigenous populations, before being completely replaced by the L4 lineage in present days, far more evidence is needed. This includes additional samples from a larger geographic distribution and additional workable skeletons from diverse archeological sites.

**Animal-Related MTBC Strains**

According to the currently available sampling and population genomics data, the animal lineages emerged from a common ancestor closely related to lineage 6 (\textit{M. africanum}) (55) (Fig. 3). Consequently, multiple mammalian host jumps occurred leading to adaptive processes and genomic erosion (66). Those animal genomes are of particular interest because genes that undergo pseudogenization or get lost are indicative of
host specificity, notably here *H. sapiens* specificity. Interestingly, three independent losses of the RD1 region have been observed in *M. microti* (67), the dassie bacillus, (68) and *Mycobacterium mungi* (69). This convergent evolution might underline the key role of the ESX-1 secretion system for infecting *H. sapiens*. Beyond the evolutionary dimension, the adaptive radiation of animal MTBC should attract more attention since animal lineages can tell us a lot about human-specific genes, which are prone to be altered or deleted in genomes belonging to the former lineages. This critical situation is illustrated by the relative paucity of published animal MTBC strains, with the notable exception of *M. bovis*, where veterinary and socioeconomic factors prevail. The few available genomes cover the following members of MTBC, *M. suricattae*, the chimpanzee bacillus, *M. microti*, *M. pinnipedi*, *M. bovis*, and *M. caprae*, but no phylogeny including all these members has been published so far. The likelihood that other ignored animal lineages exist in the field is high; a good hint would be to further investigate in the direction of social or highly promiscuous mammal species where the settlement of epidemic episodes are favored.

**Zooming into the Lineages**

One of the major advances linked with WGS is the possibility to switch to population genetic approaches in the field of *M. tuberculosis* since enough SNPs can be accumulated in the evolutionary history and the coalescence of local populations. For instance, up to 0.4 mutations per genome per year can be accumulated (70). Applying such an approach, Comas et al. (49) scrutinized the evolutionary history of the Beijing lineage, an important member of lineage 2. The Beijing family attracted much attention because its members are hypervirulent in mouse models, spread quickly in Eurasia and Western Europe, and are associated with multidrug resistance (71). The family TMRCA was estimated at 8,000 years coinciding with the rise of agriculture in the Yangtze River region, the domestication of crops and the onset of Chinese farmer populations. Interestingly, the Bayesian skyline of the Beijing family matched pretty well the one obtained from the human mitochondrial haplogroups from East Asia, confirming this likely scenario. Moreover, the dating of the Beijing family is relatively congruent with former analyses based on MIRU typing (38), but also with more recent data obtained from a large collection of 5,000 Beijing strains (72). In the later publication, Merker et al. (72) unraveled the genetic structure and global spread of the Beijing lineage; this lineage is globally distributed but still entails a fine-scale genetic structuring. The authors detected six clonal complexes (CCs) and one basal lineage; those CCs proved to be strongly associated with geographical entities (Fig. 4). They also confirmed that this lineage initially originated in the Far East 6,600 years ago from where it radiated worldwide in several waves. This was illustrated by a negative correlation ($r^2 = 0.626$) between the mean allelic richness of the strains and their distance from the Yangtze River. An ancestral East Asian population of strains, mostly endemic, that gave rise to new variants following different migration routes, can explain this pattern. The consequence of this scenario is a stepping-stone propagation of the germs, followed by successive bottlenecks, resulting in genetic erosion with increasing distance from the source. The situation is similar for *H. sapiens* and its little companion *Helicobacter pylori*, where the highest genetic diversity can be observed in Africa and the lowest one in South America (73, 74). Worth mentioning are the contrasted profiles between the ancestral strains (CC6 and BL7) that only marginally dispersed from their area of endemicity and the other derived CCs that successfully spread at continental scales. CC5 is probably the best illustration to show how a minor variant, originating from Southeast Asia, spread some 1,500 years ago into the Pacific and increased its frequency due to drift and successive founder effects, culminating at a more than 90% prevalence in Micronesia and Polynesia. One of the most striking features was the evolutionary history of CC1, also called the central Asian clade, and CC2, the Russian clone. The first CC spread westward, becoming highly predominant in central Asia and around the Black Sea, and the latter one became predominant in Russia and Eastern Europe. Both CCs had the highest clustering rates for MDR strains, indicating population expansion amplified by the recent transmission of MDR strains. The demographic success of CC1 and CC2 was confirmed by coalescent-based analyses, and their expansion dated back some 200 to 250 years ago.

These recent expansions remarkably match known episodes of migration in Asia. Indeed, several waves of Chinese refugees migrated to the Russian empire, especially Kyrgyzstan, Kazakhstan, and Uzbekistan, as a consequence of a series of national uprisings from 1861 to 1877, which might have driven the expansion of the CC1 and CC2 strains in these regions (75). These recent western expansions are probably superimposed on a more historical, continuous flux of the different Beijing sublineages westward along the Silk Road. After a tip-dating calibration, the authors reconstructed the demogenetical changes in the Beijing lineage based on a
subset of 110 genomes, and they detected a two-step increase in *M. tuberculosis* population size. The first expansion corresponded with the industrial revolution and the second one took place at the end of the 19th century, fitting the information gathered from historical and medical records. This trend was more pronounced for the most-westward distributed clonal complexes and the combined epidemic growth periods resulted in a
10-fold increase of *M. tuberculosis* Beijing’s effective population size. The mild population drop that followed the expansion phase took place in the early 1960s and might be linked to the democratization of the antibiotic use. The analysis also captured a last tiny population growth that matches the rise of the HIV epidemics. To conclude, this study demonstrated the power of NGS to explain past and yet “uncaptured” migratory paths from a single lineage and how societal changes impact tuberculosis demography and epidemics. The sudden success of some lineages can even result in full lineage replacements, as exemplified in other pathogens such as *Salmonella enterica* serovar Typhi (76), highlighting the need for and power of population genomics. In the same year, Luo et al. (77) analyzed a novel data set comprising whole-genome sequences of 358 East Asian strains belonging to the Beijing family. The authors applied the molecular clock of Comas et al. (49), which is much faster than the mutation rate implemented by Merker et al. (72). Consequently, both scientific teams reported a demographic expansion of the Beijing family, similar genetic structure and spread, but strongly disagreed on the timing and TMRCA calculations. This situation might be disturbing for the nonspecialist, and it deserves specific explanations. The molecular clock issue will be addressed in greater detail in “The relativity of the clock” (see below), which might help to clarify the situation and propose analytical improvements.

Another lineage that attracted much attention is lineage 4 (the Euro-American lineage) that circulates in Aboriginal and French Canadian communities (78). Some sublineages were introduced in the indigenous populations in the mid-18th century and spread westward through canoe routes until 1850, illustrating the impact of recent trans-Atlantic migrations on remote North American communities. It is particularly worrying to see that the Inuit living in the Nunavik region of Québec present an incidence 50-fold higher than the Canadian average. In a recent population genomics analysis, Lee et al. (79) disentangled the genetic diversity and population structure of 163 *M. tuberculosis* strains scattered in 11 remote Inuit villages. Their main finding confirmed that all patients harbored either one of two sublineages belonging to lineage 4; the TMRCA of the main sublineage, represented by 94% of the strains, dated back to the early 20th century. This result shows that the spread of tuberculosis was not interrupted after the fur trade decline, but that indigenous communities are still prone to “foreign”-mediated epidemics.

If we focus at microevolutionary scales, we reach the borders of the molecular epidemiology field and identification of transmission chains (80, 81). Lee et al. (79) nicely showed that pairs of isolates within villages had significantly fewer SNPs than pairs from different villages (6 versus 47), hinting toward intravillage chain transmissions.

**Toward a Universal Taxonomic Nomenclature?**

One of the major difficulties that a nonspecialist faces when he or she goes through the tuberculosis literature is the fluctuating and evolving nomenclature concerning the different lineages (see Table 1). The nomenclature was mainly driven by a couple of leading teams, starting from phage typing (82), regions of difference parsimony analyses (3), MIRU cladograms (35, 38), extended MLST trees (83), SNP sets (41, 84, 85), and ultimately whole-genome-based phylogenies. With the drop of the costs of Illumina and PacBio sequencing, thousands of new genomes became available, leading to the discovery of fine-scale phylogenetic structuring but also to the unearthing of new lineages. To facilitate the navigation in this growing complexity, Coll et al. (86) proposed a novel SNP-based bar code approach and implemented the PhyTB tool related to the PhyloTrack library. This numeric code relies on a subset of 62 canonical SNPs gathered from essential genes under negative selection that resolves all seven lineages and another 55 sublineages (Fig. 5). This approach can be upgraded and can evolve with the ongoing sequencing effort.

**The Relativity of the Clock**

**Substitution Rate Estimates**

Deciphering the evolutionary history of tuberculosis is highly dependent on a rigorous estimation of the molecular clock. One effective way to estimate the substitution rate is to focus on recent epidemics linked to a clone (70), retrospective observational studies (80, 87, 88), or even better on measuring the pace of mutational events within a host (89). The concept behind such approaches is that *M. tuberculosis* is composed of “measurably evolving populations” (90–92), meaning that whole genomes accumulate novel mutations over time frames of months to years. Convincingly, all these WGS studies reported congruent estimates of 0.3 to 0.5 SNP per genome per year, which translates roughly into $1 \times 10^{-7}$ substitutions per nucleotide per year, with no notable difference between hosts (human or macaque). This mutation rate places *M. tuberculosis* at the lower bound of bacterial species, compared with *Staphylococcus aureus* displaying a mutation rate of 1 to $2 \times 10^{-6}$.
Escherichia coli of $5 \times 10^{-6}$ \cite{95,96}, and the mismatch repair-lacking $H. pylori$ of $1 \times 10^{-5}$ \cite{97,98} to $7 \times 10^{-4}$ substitution per nucleotide per year during the acute phase of infection \cite{99}. Another approach to calibrate the clock relies on the high similarity of the human mtDNA-based phylogenies and the MTBC human-specific phylogeny, anchoring the Southeast Asian Oceanian lineage 1 with the human macrohaplogroup M \cite{49}. This alternative strategy resulted in a substitution rate estimate of $2.58 \times 10^{-9}$ substitutions per site per year. These two substitution rates are rather incompatible and divergent. So the question is, how can they be combined into a single model?

A way to present the problem is to invoke the fields of quantum physics and relativity to build a couple of metaphors. For example, the observer effect and the Heisenberg uncertainty principle stipulate that there is trade-off in capturing simultaneously the position and the momentum of a particle, meaning that obtaining the exact position will lower the information concerning the momentum. In the same way, applying a short-term mutation rate to a $M. tuberculosis$ data set covering...
a large temporal scale will likely provide reliable information concerning the terminal nodes and the demographic changes in the past century, but will perform poorly in terms of TMRCA and vice versa. The other metaphor concerns the theory of relativity where the faster the relative velocity is, the greater the magnitude of time dilatation will be. Again, here we can imagine that the substitution rate is a function of time and may vary following a yet-to-discover mathematical law.

These concepts have indeed some biological meaning, as we shall see. The substitution rate refers to the rate at which nucleotide changes become fixed in populations. This notion differs from the mutation rate, i.e., the rate at which novel mutations arise. In the latter case, some slightly deleterious mutations will be progressively removed by purifying selection, gradually in large populations and more stochastically in small ones. Accordingly, the mutation rate corresponds to the upper limit of mutational changes acquired per unit of time in a given biological system (100). Therefore, the combination of selective pressure (purifying selection), possible saturation at variable sites, and demographic fluctuations will shape the time dependency of evolutionary rates. This trend was noticed based on strong discrepancies between molecular and paleontological dating (101), reviewed by Ho et al. (102), but also subject to some controversies (103). However, there is growing empirical evidence for an exponential-decay law of the substitution rate, fluctuating between two natural boundaries, the mutation and the long-term substitution rates, as exemplified in New Zealand fish species (104), birds and primates (105, 106), and *Vibrio cholerae* (107). This pattern might also be more effective and important in the relative short term (over centuries) for bacterial species and viruses, since they possess much shorter generation times than, e.g., large vertebrates. This is exemplified in Fig. 6, where a strong negative linear correlation between the evolutionary rate and time (both log-transformed) could be detected in three bacterial species, based on available complete genomes.

Consequently, there is an urgent need to reunify the analyses obtained from different scientific teams.

---

**FIGURE 6** Consistent with a general pattern for measurably evolving populations, the evolutionary rates of microbial pathogens decrease as a function of the time span over which they are estimated. Data shown are selected representative examples, including one group of RNA viruses and several bacterial pathogens. Adapted from Biek et al. (100).
implementing contrasting short- or long-term substitution rates on *M. tuberculosis* data sets (108) under a same mathematical model. To reach this goal, we will have to define the parameters that define the J-shaped curve of the etiological agent of tuberculosis and to develop, and extend, tools like BEAST that will integrate in a unique coalescent framework, a substitution rate dependent on time (109). However, this task remains challenging, since there is a paucity of reliable calibration points for the intermediate time frames.

**Other Limitations**

Some additional features must be mentioned that possibly complicate the clock estimates. Among them, we have to consider both intrachromosomal and interlineage variations in the ticking rates. The 4.4 Mb *M. tuberculosis* genome presents highly variable repetitive genetic regions, encompassing genes such as the PPE, PE_PGRS, and ESX families. Those gene families are prone to increased mutation rates, are difficult to assemble, and are often removed from the analyses. Therefore, the accuracy of the trimming step might explain some outliers observed in terms of mutation rate estimates. Furthermore, Martincorena et al. (110) detected mutational hot and cold spots across 2,659 genes from a collection of 34 *E. coli* strains. Lower rates were observed in highly expressed genes and there is no strong argument that *M. tuberculosis* should behave differently. At shorter timescales, mutations affecting genes involved in MTBC DNA repair (111) can inflate the mutation rate, resulting in hypermutator phenotypes. These transient phenotypes are adaptively cardinal and are involved in the fast acquisition of SNPs conferring resistance to second-generation antibiotics. For example, one copy of the major replicative DNA polymerase, dnaE2, proved to be a mediator of *M. tuberculosis* survival through inducible mutagenesis and contributed to the emergence of drug resistance *in vivo* (112). At a higher hierarchical level, there is growing evidence that some Beijing sublineages undergo faster evolutionary rates relative to other human tuberculosis lineages. Ebrahimi-Rad et al. (113) detected alterations in *mut* genes, mostly missense mutations that improve the adaptability of the W-Beijing clade. In the same vein, Ford et al. (114) demonstrated that cultured *M. tuberculosis* strains from lineage 2 acquired *in vitro* drug resistance against isoniazid and ethambutol three times faster than *M. tuberculosis* strains from lineage 4, invoking again contrasted mutation rates between lineages. However, this difference remains relatively modest compared with *E. coli* (115) and *Pseudomonas aeruginosa* (116), where mutator phenotypes are orders of magnitude more mutable than the wild-type strains. The latter species harbor a mismatch-repair system whose dysfunction increases mutation and recombination rates; in that respect, they differ from *M. tuberculosis*, which is lacking such a system (117).

A final complication can be called up, linked to the peculiar strategy employed by the tubercle bacillus to survive in the host. During latent infection, which accounts for most of its life history, *M. tuberculosis* is in a dormant state with little to no replication activity. This latent stage dramatically contrasts with the active stage of the disease in which clonal multiplication occurs in the lung. The question can therefore be raised whether mutation rates differ between the two different stages and whether this might impact the evolutionary history of the MTBC and the TMRCA estimates. This problem was targeted by a team at the Harvard School of Public Health (89); the authors were able to show, based on whole-genome sequencing of *M. tuberculosis* isolated from cynomolgus macaques, that the mutation rates were similar during latent and active disease. Since most of the chromosome mutations appear during the replication process, this result was somehow unexpected. One of the explanations proposed relies on the high oxidative stress that the bacilli undergo in the macrophage phagolysosome resulting in cytosine deamination or the formation of 8-oxoguanine (118, 119). Alternatively, *M. tuberculosis* keeps dividing more actively during latency than previously thought.

**PERSPECTIVES**

High financial investments coupled with novel methodological and analytical approaches have driven human population genomics to the upper edge of scientific excellence. The study of the natural host of *M. tuberculosis* reached a new summit with the rise of the study of ancient DNA (120). In 2006, Svante Pääbo and his colleagues analyzed 1 million bp of a 38,000-year-old Neanderthal fossil using the 454 technology (121) and were able to handle the small degraded DNA molecules. Among the most common “lesions,” depurination and deaminated cytosine residues near the fragment ends were reported (122). Four years later, thanks to a methodological switch toward Illumina sequencing, they provided the scientific community with the first draft genome of a Neanderthal gathered from three individuals that lived in the Vindija cave in Croatia (123). This major technological step, accompanied with suites of bioinformatics tools, opened the door to late Pleistocene
genomics (124–126) and allowed the discovery of new Hominin lineages, like the Denisovan (127), an extinct relative of Neanderthals.

The first bacterial paleogenomes came just after, notably with the first draft genome of Y. pestis, the etiological agent of the plague (128). This ancient genome at an average 30-fold coverage was isolated using targeted capture from Black Death victims in London who died in approximately 1348 to 1350. Even more impressive was the publication of a Bronze Age Y. pestis genome from Asia, 3 millennia earlier than any historical records of plague (129). Auspiciously, the bacteria belonging to the MTBC present a complex waxy and hydrophobic cell wall, facilitating therefore the conservation of the DNA molecules. With the development of the field of paleomicrobiology we might soon have the unique opportunity to sample temporal series of skeletons showing stigmata of tuberculosis infection (up to 5% of the patients). Thanks to precise radiocarbon dating, we might even capture from Black Death victims in London who died in the approximate year 1350. Even more impressive is that from the late Pleistocene to the present (17, 130–133). The accumulation of heterochronous M. tuberculosis paleopopulation genomics will ultimately solve the molecular clock issue and possibly reunify the fields of evolutionary mycobacteriology, paleoepidemiology, and paleopathology. The first steps have been already accomplished, with 100–562, 250–63, and 1,000-year-old 35 genomes made available.

Beyond this rather technical issue, the spread of the disease lines through time, lineage replacements, extinct lineages discovery, and eventually a complete revision of current models will be driven by paleomicrobiology. Once Pandora’s box is open, we might consider the possibility that Neanderthals may have faced M. tuberculosis. If true, two conflicting scenarios can be tested. Neanderthals harbored their own MTBC-like ancestor or they contracted a H. sapiens-associated strain during their coexistence with modern humans, some 30,000 years ago in southwestern Europe. Last but not least, ancestral genomes will also improve our understanding of pathogen adaptation and the pace and dynamics of coevolution with its natural host. Studying tuberculosis remains a challenging task, but for sure, we live in interesting times.

ACKNOWLEDGMENTS
We gratefully acknowledge the contributions and comments of Jean-Philippe Rasigade, Olivier Dutour, and Jan Willem Dogger on the manuscript. We also thank the Ecole Pratiques des Hautes Etudes for financial support via the priority research action Grant (ARP).

REFERENCES

[16] ASMscience.org/MicrobiolSpectrum


Evolutionary History, Demography, and Spread of MTBC

Beijing lineage. *Nat Genet* 47:242–249 [http://dx.doi.org/10.1038/ng.3195].


