Epidemics 51 (2025) 100821

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Onset of infectiousness explains differences in transmissibility across *Mycobacterium tuberculosis* lineages

Etthel M. Windels [a,b,c,*], Cecilia Valenzuela Agüí [a,c], Bouke C. de Jong [d], Conor J. Meehan [d,e], Chloé Loiseau [b,f], Galo A. Goig [b,f], Michaela Zwyer [b,f], Sonia Borrell [b,f], Daniela Brites [b,f,1], Sebastien Gagneux [b,f,1], Tanja Stadler [a,c,*,1]

[a] ETH Zürich, Basel, Switzerland
[b] Swiss Tropical and Public Health Institute, Allschwil, Switzerland
[c] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[d] Institute for Tropical Medicine, Antwerp, Belgium
[e] Nottingham Trent University, Nottingham, UK
[f] University of Basel, Basel, Switzerland

ARTICLE INFO

Keywords:
Phylodynamics
Mycobacterium tuberculosis
Transmission

ABSTRACT

Mycobacterium tuberculosis complex (MTBC) lineages show substantial variability in virulence, but the epidemiological consequences of this variability have not been studied in detail. Here, we aimed for a lineage-specific epidemiological characterization by applying phylodynamic models to genomic data from different countries, representing the most abundant MTBC lineages. Our results suggest that all lineages are associated with similar durations and levels of infectiousness, resulting in similar reproductive numbers. However, L1 and L6 are associated with a delayed onset of infectiousness, leading to longer periods between subsequent transmission events. Together, our findings highlight the role of MTBC genetic diversity in tuberculosis disease progression and transmission.

1. Introduction

Human tuberculosis (TB) is characterized by a large heterogeneity in clinical and epidemiological features (Coscolla and Gagneux, 2010; Cadena et al., 2017). Although several host and environmental factors partially underlie this variability, there is increasing evidence that genetic diversity within the *M. tuberculosis* complex (MTBC) also plays a role in TB disease presentation and transmission dynamics. Ten human-adapted MTBC lineages (L1 to L10) have been identified to date (reviewed in Orgeur et al., 2024; Guyeux et al., 2024). L1 to L6 are globally the most abundant lineages, where L1, L5, and L6 are often called the "ancient" lineages, and L2, L3, and L4 are commonly referred to as the "modern" lineages (Brosch et al., 2002; Gagneux, 2018; Bottai et al., 2020).

Several animal and macrophage infection studies have shown reduced virulence of strains belonging to "ancient" lineages compared to strains from "modern" lineages, observed as restricted *in vivo* growth, host immune modulation, and disease severity (reviewed in Coscolla and Gagneux, 2010; Coscolla and Gagneux, 2014; Tientcheu et al., 2017; Peters et al., 2020). Consistent with this reduced virulence, molecular epidemiological studies have reported a lower transmissibility of "ancient" compared to "modern" lineages (Yang et al., 2012; Guerra-Assunção et al., 2015; Asare et al., 2018; Holt et al., 2018; Sobkowiak et al., 2020; Freschi et al., 2021; Zwyer et al., 2023; Gröschel et al., 2024), with the majority of these studies using clustering rates and/or terminal branch lengths (TBLs) to quantify transmission. These metrics indirectly estimate the time between subsequent transmission events, but do not explicitly consider patient infectiousness (i.e. the ability to transmit) during that time, resulting in an incomplete picture of the transmission dynamics for two reasons. First, the time between transmission events also includes the time between infection and the onset of infectiousness, and is hence not necessarily a measure for how rapidly infectious patients spread the disease. Second, the duration of infectiousness directly affects the effective reproductive number (R_e), i.

* Corresponding authors at: ETH Zürich, Basel, Switzerland.
E-mail addresses: etthel.windels@bsse.ethz.ch (E.M. Windels), tanja.stadler@bsse.ethz.ch (T. Stadler).
[1] Equal contribution.

https://doi.org/10.1016/j.epidem.2025.100821
Received 6 September 2024; Received in revised form 30 January 2025; Accepted 7 March 2025
Available online 11 March 2025

e. the expected number of secondary cases caused by a single infected individual, which determines whether case numbers increase or decrease over time. An additional limitation of clustering rates and TBLs is that they ignore (potentially lineage-specific) variation in clock rate and sampling intensity, which both affect the genetic distance between sampled isolates (Menardo, 2022). A better resolved and more accurate description of patient infectiousness and transmissibility for the different MTBC lineages is indispensable to anticipate the future epidemic spread and prevalence of these lineages.

Phylodynamic birth-death models explicitly model the processes of molecular evolution, transmission, becoming (non-)infectious, and sampling (Stadler, 2009; Stadler, 2010; Kühnert et al., 2016), which allows disentangling the relevant evolutionary and epidemiological characteristics. Here, we used different versions of the birth-death model to characterize the epidemiological features of the main MTBC lineages in unprecedented detail. As we aimed for a comprehensive epidemiological comparison of lineages, the models were applied to genomic data from four different sampling locations where "ancient" (L1 and L6) and "modern" (L2, L3, and L4) lineages co-circulate (Malawi, Tanzania, The Gambia, and Vietnam). Our results suggest that all lineages (except for L1 in Vietnam) are characterized by an R_e close to 1, and are therefore on average not expected to change much in relative abundance over time. However, L1 and L6 are characterized by longer periods between subsequent infection events than L2, L3, and L4. Our data suggest that this is due to a delayed onset of infectiousness in L1 and L6-infected individuals, rather than an overall reduced level of infectiousness. These findings provide new insights into the implications of MTBC genetic diversity for transmission and disease progression.

2. Results

We used publicly available whole-genome sequencing data from four countries where "ancient" and "modern" lineages co-circulate: 1684 sequences from Malawi (L1, L2, L3, and L4) (Guerra-Assunção et al., 2015), 921 sequences from Tanzania (L1, L2, L3, and L4) (Zwyer et al., 2023), 1086 sequences from The Gambia (L2, L4, and L6) (Gehre et al., 2024), and 1623 sequences from Vietnam (L1, L2, and L4) (Holt et al., 2018) (see Materials and Methods for details on the study populations). Each of these settings is characterized by a low prevalence of multidrug resistance (MDR; see Materials and Methods). Simple metrics such as clustering rates (Fig. S1a) and TBLs (Fig. S1b) for these populations suggest that transmissibility is lowest for L1/L6 and highest for L2, largely in accordance with previous studies (Yang et al., 2012; Guerra-Assunção et al., 2015; Asare et al., 2018; Holt et al., 2018; Sobkowiak et al., 2020; Freschi et al., 2021; Zwyer et al., 2023; Gröschel et al., 2024).

By applying phylodynamic birth-death models to these genomic data, we aimed for a more detailed epidemiological characterization of the different MTBC lineages, including their associated onset and duration of infectiousness, level of infectiousness (number of secondary transmissions per infected individual per unit of time), and the resulting reproductive number. As the classical birth-death model (also called "single-type birth-death model") does not distinguish between non-infectious and infectious infected individuals, we used an extension based on the multi-type birth-death model, allowing for an initial non-infectious period in infected individuals (Fig. 1). This is achieved by introducing an epidemiological compartment representing infected non-infectious individuals in the population. The non-infectious/infectious dichotomization in this model is purely based on the ability to cause secondary infections, rather than on symptoms or radiographic/microbiological markers. This simplifies the continuous clinical disease spectrum characterizing TB infection and disease (Drain et al., 2018; Kendall et al., 2021), but potentially represents a major improvement in accuracy compared to the single-type birth-death model as well as less fine-grained transmissibility metrics. This multi-type birth-death model was fitted to the genomic data from Malawi, Tanzania, The Gambia, and Vietnam, with priors reported in Table S1 and visualized in Fig. S2 (see Materials and Methods for details on the model). Except for L1 in Vietnam, the posterior estimates for the R_e were close to 1 for all lineages in all sampling locations (Fig. 2a; Table S2). This suggests that in

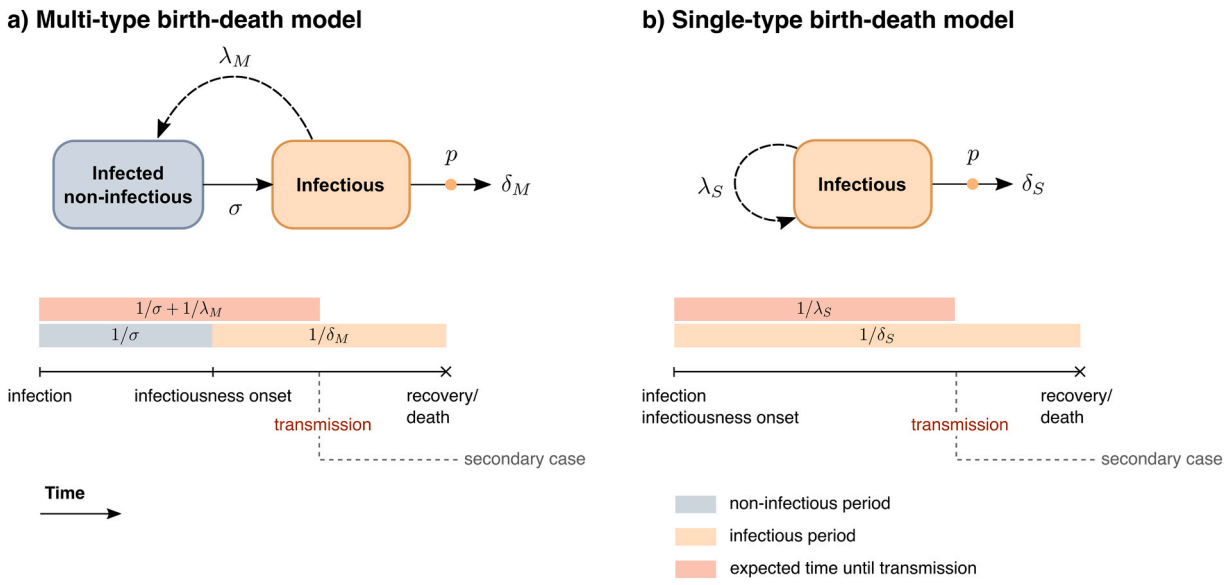

Fig. 1. Schematic representation of the phylodynamic birth-death models used in this study. a) The multi-type birth-death model distinguishes individuals who are infected but not yet infectious from individuals who are infectious and generate secondary infections at rate λ_M. A new infection results in a new individual in the "infected non-infectious" compartment, who moves to the "infectious" compartment at rate σ. Hence, the expected duration of the non-infectious period equals $1/\sigma$. Due to this initial non-infectious period, $1/\lambda_M$ represents the expected time until secondary infection since the start of the infectious period. Individuals become non-infectious through recovery or death at rate δ_M, resulting in an average infectious period of $1/\delta_M$. b) The classical, single-type birth-death model is a more constrained version of a) including only a single compartment of infected individuals. Upon infection, occurring at a constant transmission rate λ_S, individuals instantaneously become infectious. Hence, the expected time between infection and the first transmission event equals $1/\lambda_S$. Individuals become non-infectious through recovery or death at rate δ_S. Consequently, the average infectious period in this model equals $1/\delta_S$ and corresponds to the total duration of infection. In both models, patients are sampled at rate p.

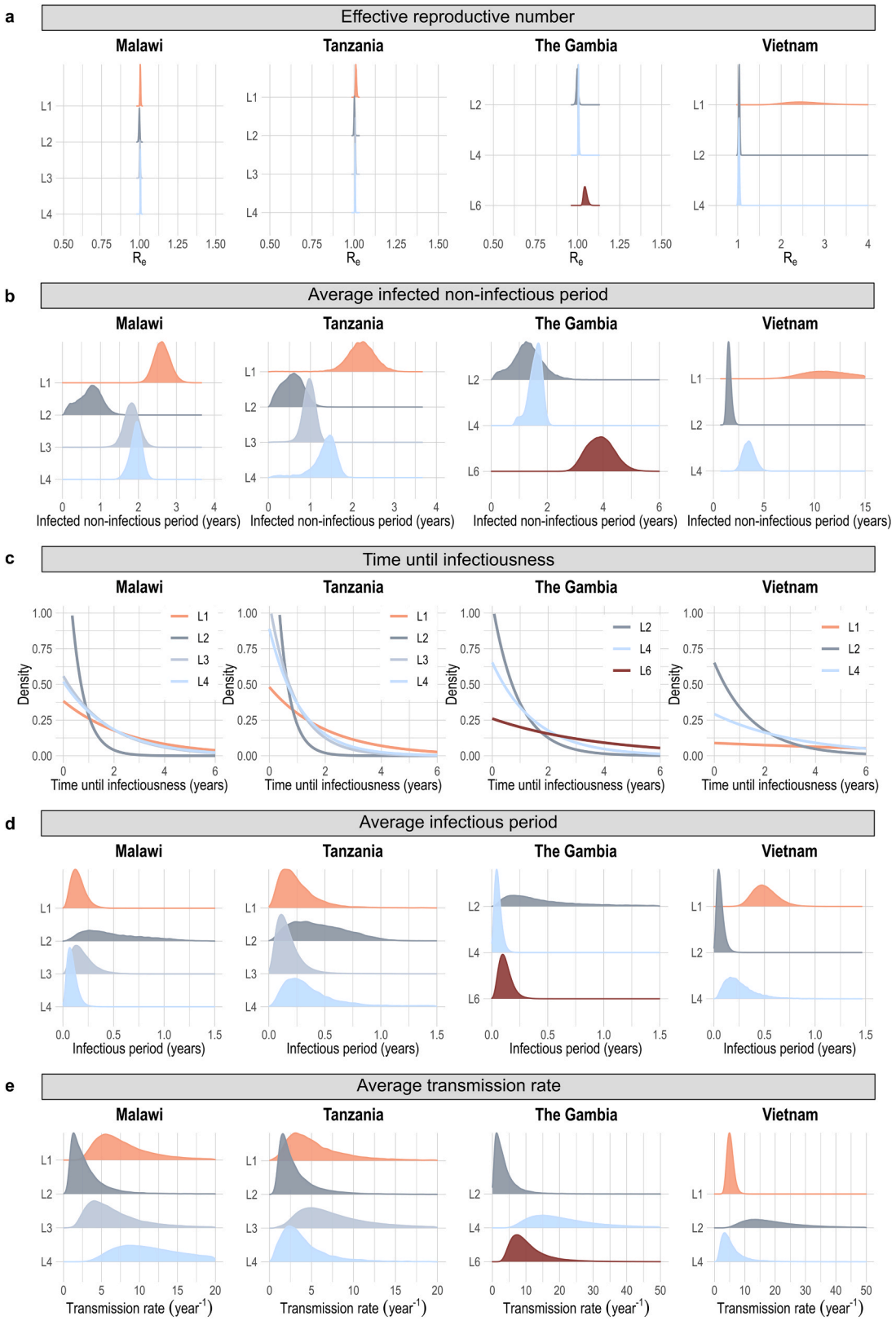

Fig. 2. Posterior estimates for the multi-type birth-death model fitted onto genomic data from different sampling locations. a) Posterior distributions of the effective reproductive number (R_e), showing estimates close to 1 for all lineages in all sampling locations (except for L1 in Vietnam). b) Posterior distributions of the average duration of the initial non-infectious period, showing the highest estimates for L1 and L6 in all sampling locations. c) Population distributions of the time until infectiousness, assuming an exponential distribution with rate parameter corresponding to the posterior mean in b). d) Posterior distributions of the average duration of the infectious period. e) Posterior distributions of the average transmission rate during the infectious period.

these countries, the relative abundance of each lineage is on average not changing considerably over time. However, the average time between infection and the onset of infectiousness was consistently estimated to be longest for L1 and L6 (Fig. 2b), with the 95 % highest posterior density (HPD) intervals for these lineages not overlapping with the other lineages (Table S3). These posterior estimates are estimates of the population average, and can be translated into population distributions of times until onset of infectiousness (Fig. 2c), showing that on average, 26 % (L1), 73 % (L2), 54 % (L3), 43 % (L4), and 23 % (L6) of infected individuals were estimated to progress to infectious TB within one year of infection (Table 1). In contrast to this lineage-specific initial non-infectious period, HPD intervals for the subsequent infectious period did overlap for the different lineages (Fig. 2d; Table S4). Although the posterior uncertainty for these parameter estimates was relatively high, especially for the smaller L2 datasets from Malawi, Tanzania and The Gambia (Fig. S2c), the data overall did not support differences between lineages in the infectious period and transmission rate (i.e. the rate of secondary case generation per infected individual) during this infectious period (Fig. 2d-e). The exception to this was L1 in Vietnam, for which the long infectious period resulted in a high R_e estimate. Together, these results suggest that all lineages are characterized by a similar duration and level of infectiousness, but that "ancient" lineages L1 and L6 are associated with a longer period of initial non-infectiousness, and hence longer periods between infections and secondary transmission events.

To assess the robustness of our results to prior assumptions, we repeated the analyses assuming different levels of underreporting (see Materials and Methods). This resulted in different absolute values for the posterior estimates, but similar relative differences between lineages (Fig. S3, Fig. S4). Moreover, we also allowed for some rates to change 30 years before the most recent sample, to rule out the influence of independent MTBC introduction events that potentially shape the early parts of the phylogenetic trees (see Materials and Methods). To limit the model complexity for these analyses, we fixed R_e to 1 for both time intervals. The parameter estimates for the most recent time interval (Fig. S5) were similar to the estimates from the main analyses (Fig. 2), suggesting limited biases due to independent introductions. Finally, since the priors in the main analyses put an infinitely small weight on very short non-infectious periods (Table S1), we reparametrized the model and tested the effect of a stronger prior support for short non-infectious periods (see Materials and Methods). Again, the relative differences between lineages remained unchanged (Fig. S6).

We further investigated the importance of the non-infectious period in explaining the genomic data by comparing the multi-type birth-death estimates to the estimates from the single-type birth-death model, where infected individuals are assumed to instantaneously become infectious (Fig. 1b; Table S5). For the single-type birth-death model, the estimated R_e was again close to 1 for all lineages at all locations under study, except for L1 in Vietnam (Fig. S7). The estimates for the expected time until secondary transmission were also similar for both models, suggesting the longest times for L1 and L6 (Fig. S8; Table 2). This average time until transmission is related to commonly used transmission metrics like clustering rates and TBLs. Similarly, the estimates for the average total infected period were in good accordance for both models, indicating robustness to the choice of model (Fig. S9; Table 2). As expected, the major difference between the two models was that the estimates from the single-type birth-death model suggest a relatively long infectious period associated with a relatively low transmission rate, while the results from the multi-type birth-death model suggest a non-negligible non-infectious period, followed by a relatively short infectious period associated with a high transmission rate (Fig. S10; Fig. S11; Table 2). To investigate which of these two models best explains the data, we performed a model selection analysis with 50 % prior weight on each model (see Materials and Methods). This analysis resulted in a clear posterior preference (83–100 % posterior support) for the multi-type birth-death model for all lineages except L2 (Table 3), which was not surprising given the consistently short non-infectious period estimated for this lineage (Fig. 2b). Together, these results further support that the longer time between transmission events for "ancient" lineages, as observed in this and previous studies, stems from a later onset, rather than a lower level of infectiousness.

The estimated branch lengths in the phylogenetic tree, informing our epidemiological parameter estimates, depend on the clock rate (number of substitutions per site per year) which is used to convert the observed genetic changes into time. In our initial analyses, we estimated the clock rate from the data, setting a relatively informative prior and allowing for lineage-specific clock rate estimates (see Materials and Methods; Table S6). However, estimates for the clock rate and time between transmissions might be confounded due to the limited clock signal in the data. To investigate the impact of this on our results, we fixed the lineage-specific clock rates to a set of values and examined the order of magnitude of difference required to explain the observed differences in branch lengths, from which the differences in the non-infectious period duration are derived. We focused on L1 and L2 in Tanzania, as these showed strong differences in the estimated non-infectious period (Fig. 2b). The results show that the L1 clock rate would need to be 16–64-fold higher than the L2 clock rate in order for the posterior distributions of these estimates to overlap (Fig. 3). This estimated difference is much larger than the two-fold difference reported before

Table 2
Arithmetic expressions used to calculate the posterior distributions of parameters that were not directly part of the model.

Parameter	Multi-type birth-death model	Single-type birth-death model
Transmission rate during infectiousness	$R_{e,M}\delta_M$	$R_{e,S}\delta_S$
Time between start of infection and first transmission event	$1/\sigma + 1/\lambda_M$	$1/\lambda_S$
Non-infectious period	$1/\sigma$	0
Infectious period	$1/\delta_M$	$1/\delta_S$
Total infected period	$1/\sigma + 1/\delta_M$	$1/\delta_S$

Table 3
Posterior probability of the multi-type birth-death model, assessed in a model selection analysis assuming 50 % prior probability of the single-type (0) and multi-type (1) birth-death model.

	Lineage 1	Lineage 2	Lineage 3	Lineage 4	Lineage 6
Malawi	1	0.48	1	1	-
Tanzania	0.98	0.35	0.95	0.83	-
The Gambia	-	0.72	-	1	1
Vietnam	1	1	-	1	-

Table 1
Posterior mean and 95 % highest posterior density interval for the estimated proportion of infected individuals who progress to infectious TB within one year of infection.

	Lineage 1	Lineage 2	Lineage 3	Lineage 4	Lineage 6
Malawi	0.32 [0.28,0.36]	0.86 [0.46,0.99]	0.43 [0.36,0.49]	0.41 [0.36,0.46]	-
Tanzania	0.38 [0.29,0.46]	0.91 [0.52,1.00]	0.66 [0.54,0.75]	0.59 [0.41,0.78]	-
The Gambia	-	0.66 [0.25,0.93]	-	0.48 [0.38,0.60]	0.23 [0.18,0.28]
Vietnam	0.086 [0.049,0.12]	0.48 [0.39,0.56]	-	0.25 [0.19,0.32]	-

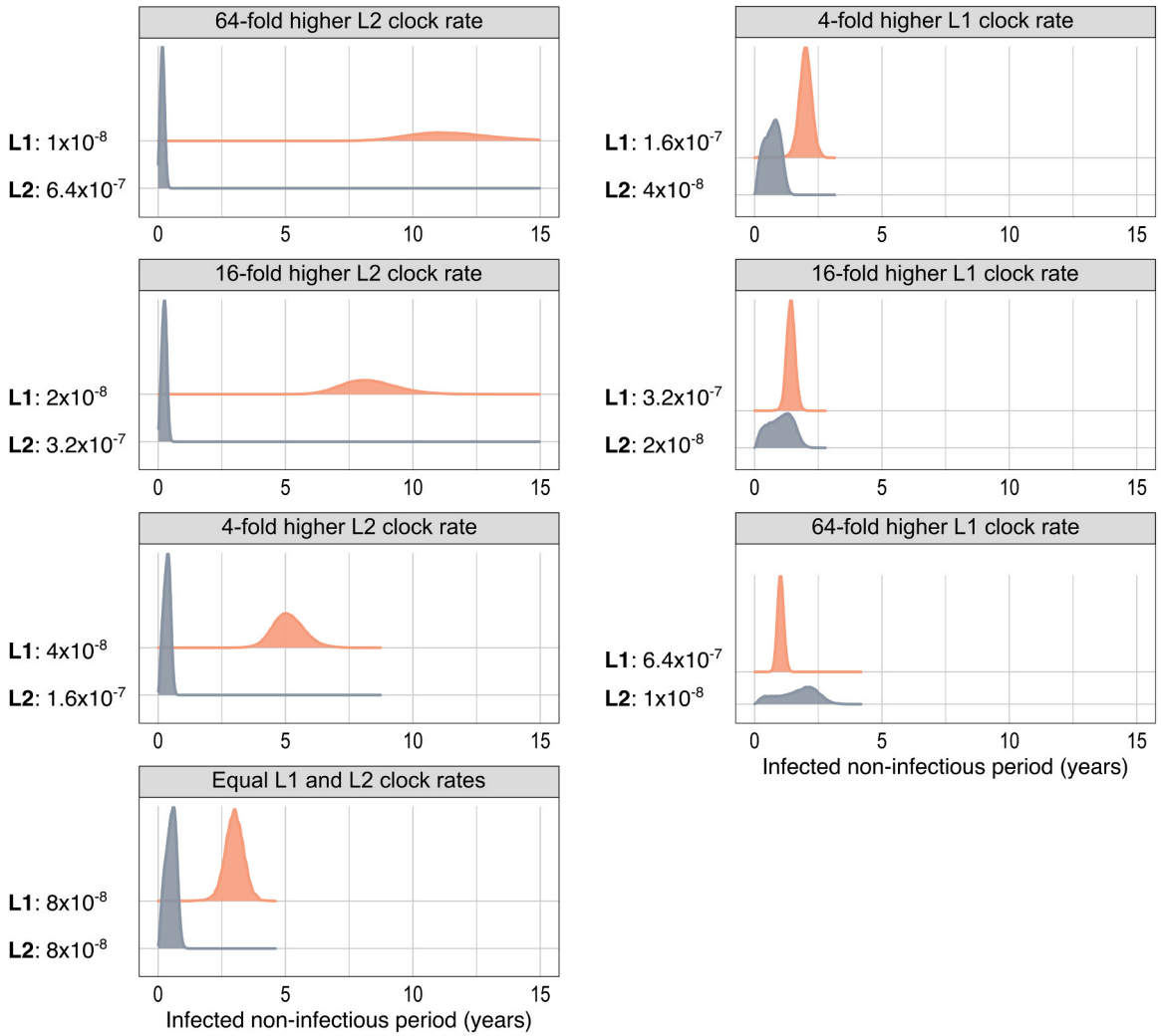

Fig. 3. Posterior distributions of the average duration of the initial non-infectious period for L1 and L2 in Tanzania, assuming different combinations of fixed clock rates (substitutions/site/year). The distributions start overlapping when the clock rate for L1 is at least 16-fold higher than for L2.

(Menardo et al., 2019) and estimated in our main analyses (Table S6). This suggests that while assumptions about the clock rate do have a significant impact on epidemiological parameter estimates, the expected clock rate differences between lineages are not sufficient to explain the observed differences in branch lengths.

3. Discussion

In this study, we applied phylodynamic birth-death models to genomic data collected in four different locations where "ancient" and "modern" MTBC lineages co-circulate, with the aim of inferring lineage-specific epidemiological characteristics. Previous studies have suggested a lower transmissibility of "ancient" lineages (mainly L1) compared to "modern" lineages (mainly L2), based on clustering rates or TBLs (Yang et al., 2012; Guerra-Assunção et al., 2015; Holt et al., 2018; Sobkowiak et al., 2020; Freschi et al., 2021; Zwyer et al., 2023; Gröschel et al., 2024), both of which indirectly measure the time between subsequent transmission events. In accordance with these previous findings, our results support a longer expected time between transmission events for L1 and L6 compared to L2, L3, and L4. By explicitly modelling (non-) infectiousness in infected individuals, we show that this difference can be robustly explained by a longer initial period of non-infectiousness in patients infected with L1 or L6 strains. Our results further suggest that this non-infectious period is followed by an infectious period for which the average duration and level of infectiousness (measured as the transmission rate, i.e. the rate of secondary case generation per infected individual) is not different between lineages, although the amount of information in the data about these parameters varies across datasets. We demonstrate that a longer initial non-infectious period for "ancient" lineages better explains the data than the alternative model of a longer infectious period starting immediately upon infection, combined with a lower transmission rate (resulting from a lower level of infectiousness) during this infectious period. Overall, our findings are in accordance with a previous household contact study showing that strains from L4 and L6 are associated with similar levels of patient infectiousness, but that L4 is associated with an increased risk of disease progression within 2 years of infection (de Jong et al., 2008).

The observed association between "ancient" lineages and slow disease progression could be linked to the stronger inflammatory immune response reported against these strains (Portevin et al., 2011; Chen et al., 2014) as well as their lower replication rate (Reiling et al., 2013; Sanoussi et al., 2017; Hiza et al., 2024), both of which have been suggested to reflect lower virulence. Another manifestation of this reduced virulence is the observed association of L1 and L6 with old age (Thwaites et al., 2008; de Jong et al., 2010; Guerra-Assunção et al., 2015) as well as with HIV co-infection (Glynn et al., 2010; de Jong et al., 2010; Guerra-Assunção et al., 2015; Asante-Poku et al., 2016), suggesting that L1 and L6 might be more likely to cause asymptomatic, non-infectious infection in individuals that are not immunosuppressed. In support of this notion, the average asymptomatic period was estimated to be longer

in countries with a high L1 burden (Ku et al., 2021), and a higher prevalence of asymptomatic TB has recently been observed in patients infected with "ancient" lineage strains (Long et al., 2024).

Our estimates of the onset of infectiousness are population averages, summarizing an underlying population distribution. Assuming that the time until onset of infectiousness is exponentially distributed (Borgdorff et al., 2011; Behr et al., 2018; Menzies et al., 2021) and that 75 % of all cases are not reported, our estimates suggest that between 23 % (L6) and 73 % (L2) of the population progresses to infectious TB within the first year of infection. These values are within the range of previous estimates (Borgdorff et al., 2011; Sloot et al., 2014; Behr et al., 2018; Emery et al., 2021; Menzies et al., 2021; Horton et al., 2023), although there is variation across countries. Assuming lower rates of underreporting resulted in lower estimated rates of disease progression. However, the exact level of underreporting is unknown and likely varies across sampling locations.

The average time between transmission events is often used as a measure for transmissibility but is not the only factor determining how rapidly the prevalence of MTBC lineages changes over time. Instead, these dynamics are also largely determined by the average number of secondary cases caused by one infected individual (R_e), which is the product of the average duration of infectiousness and the transmission rate during infectiousness. Our results show that the average R_e is consistently estimated around one for all lineages in all locations under study. The exception to this is L1 in Vietnam, which could be due to the dominance of sublineage L1.1.1, reported to demonstrate increased transmission potential (Stanley et al., 2024), whereas L1.1.2, L1.1.3, L1.2.1, and L1.2.2 are circulating in Malawi and Tanzania. However, it should be noted that the posterior uncertainty around this R_e estimate is high, which might be related to the limited temporal signal in the data as a result of the relatively short sampling period. Although our R_e estimates represent time averages, and we did not investigate changes in R_e through time, an R_e of one implies that the TB prevalence per lineage remains relatively constant over time. However, the longer period between transmission events observed for L1 and L6 implies a lower turnover rate within the population of infectious individuals.

Genetic distances between sampled isolates, and consequently also branch lengths in a phylogenetic tree, are not only determined by the rate of transmission, but also by the clock rate and the sampling density. In contrast to clustering methods (applying a fixed SNP threshold) and methods based on TBLs, which both implicitly assume equal clock rates for all lineages, the phylodynamic models used here allow for the simultaneous inference of clock rates from the genomic data. Since the clock signal in *M. tuberculosis* data is intrinsically weak, we tested different scenarios by fixing the clock rate to a set of different values. These analyses show that, in order for the clock rate only to explain the branch length differences, the clock rate for L1 would need to be 16–64-fold higher than for L2, which is considerably more than the two-fold difference estimated in a previous systematic analysis (Menardo et al., 2019).

It is worth noting that the non-infectious phase at the start of infection could be associated with reduced bacterial replication and, consequently, a lower mutation rate. While there is some evidence for mutagenesis occurring during the non-infectious phase (Ford et al., 2011; Lillebaek et al., 2016; Colangeli et al., 2020), a reduced mutation rate could bias the estimated duration of this phase, especially when only the genetic diversity is taken into account (as is done in clustering and TBL analyses). In our phylodynamic analyses, we cannot assign different rates to the different infectious stages but assume an average rate. Future methodological work on assigning such different rates will enable the quantification of potential differences.

Except for the L6 culture bias (Sanoussi et al., 2017) accounted for in the priors, we assumed no lineage-specific sampling biases. Nonetheless, the datasets used in this study mainly contain isolates from patients who presented themselves at a healthcare center, most likely following symptom development. While this includes patients who potentially went through an asymptomatic phase before progressing to active TB disease, it does not include infected individuals who never develop symptoms. The incidence of asymptomatic TB cases who never get diagnosed is currently unknown but might be higher in individuals infected with "ancient" lineage strains (Long et al., 2024). If this is indeed the case, this would imply a lineage-specific undersampling and might affect our results. Without genomic data collected from these asymptomatic cases, such biases are challenging to control for. However, these cases are only relevant for the transmission dynamics if they do not represent dead ends in transmission chains, which would imply that they are infectious despite being asymptomatic. While some studies do suggest some degree of infectiousness in asymptomatic cases (Xu et al., 2019; Frascella et al., 2021; Lau et al., 2022), more studies are needed to determine the strength and variation in infectiousness in these individuals.

Except for the differential presence of the TbD1 genomic region (Brosch et al., 2002) and some recently identified regions of difference (Behruznia et al., 2024), little is known about the genetic differences between "ancient" and "modern" lineages that may underlie the differential TB progression rate observed in this study. Furthermore, the relevance of the "ancient/modern" dichotomy can be questioned for several reasons. First, this and previous studies show clear epidemiological differences between L1/L6 and L2, but L3 and L4 seem to show intermediate behavior. Second, this classification does not properly account for the recently discovered lineages L7–10. Additionally, within-lineage diversity, especially within the most diverse L1 (Coscolla and Gagneux, 2014), might further complicate the picture. As noted above, observed differences between L1 estimates in different locations might be due to the dominance of different sublineages, emphasizing the need for a more fine-grained epidemiological characterization. This would require an unbiased sample set representing the genetic diversity in the MTBC at higher resolution.

Taken together, our results demonstrate that the MTBC lineages circulating in Malawi, Tanzania, The Gambia, and Vietnam are associated with a similar effective reproductive number, but different onset of infectiousness. In particular, the slower progression to an infectious disease state, as observed for L1 and L6, results in longer periods between transmission events. These results can explain why the prevalence per lineage tends to stay relatively stable over time, despite the higher incidence of "modern" compared to "ancient" lineages in these settings. Our findings provide insights into the epidemiological consequences of MTBC genetic diversity, but more studies are needed to narrow down the underlying genetic determinants.

4. Materials and methods

4.1. Study populations

4.1.1. Malawi

We retrieved publicly available sequences of isolates collected from adults with culture-confirmed pulmonary or extrapulmonary TB diagnosed through passive case finding at the hospital and peripheral health centers in Karonga District, northern Malawi, between 1995 and 2011 ($n = 1684$) (raw reads available in the European Nucleotide Archive under project accession numbers PRJEB2358 and PRJEB2794) (Guerra-Assunção et al., 2015). The lineage distribution in the genomic dataset is as follows: L1: $n = 266$, L2: $n = 70$, L3: $n = 188$, L4: $n = 1160$ (subsampled to $n = 400$ for computational feasibility), with 9 MDR isolates in total. The reported incidence of smear-positive TB in adults in the district during the sampling period corresponds to 87–124 cases per 100,000 people per year (Guerra-Assunção et al., 2015).

4.1.2. Tanzania

We used previously sequenced isolates ($n = 921$) from a cohort of sputum smear-positive and GeneXpert-positive adult pulmonary TB patients. These patients were prospectively recruited at the Temeke

District hospital in Dar es Salaam, Tanzania, between 2013 and 2019, in the context of the National TB and Leprosy Programme - Tanzania (raw reads available in the European Nucleotide Archive under project accession number PRJEB49562) (Zwyer et al., 2023). The lineage distribution in the genomic dataset is as follows: L1: $n = 137$, L2: $n = 74$, L3: $n = 426$ (subsampled to $n = 400$), L4: $n = 284$, with 2 MDR isolates in total. In 2020, 3994 TB cases were notified in Temeke (Jerry Hella, personal communication).

4.1.3. The Gambia

We used previously sequenced isolates ($n = 1086$) collected in the context of a cluster randomized trial (ClinicalTrials.gov Identifier: NCT01660646) conducted between 2012 and 2014 in the Greater Banjul Area, The Gambia (raw reads available in the European Nucleotide Archive under project accession number PRJEB53138). Patients in the control arm were diagnosed through passive case finding, whereas patients in the intervention arm were diagnosed through a combination of passive and enhanced case finding, although no impact of the intervention on the transmission dynamics was observed (Gehre et al., 2024). The lineage distribution in the genomic dataset is as follows: L2: $n = 35$, L4: $n = 735$ (subsampled to $n = 400$), L6: $n = 316$, with 10 MDR isolates in total. The reported TB incidence in The Gambia during the sampling period corresponds to 176 per 100,000 people per year (World Bank, 2024).

4.1.4. Vietnam

Raw reads were retrieved from the NCBI BioProject database (accession ID: PRJNA355614). These were obtained from adults with smear-positive pulmonary TB diagnosed through passive case finding at eight district tuberculosis units in Ho Chi Minh City, Vietnam, between 2008 and 2011 ($n = 1623$) (Holt et al., 2018). The lineage distribution in the genomic dataset is as follows: L1: $n = 380$, L2: $n = 1053$ (subsampled to $n = 400$), L4: $n = 190$, with 64 MDR isolates in total. The reported annual incidence of pulmonary TB in Ho Chi Minh City during the sampling period corresponds to ~11,000 cases (Holt et al., 2018). Sample collection dates were kindly provided by the authors of the study. Since only sampling year and month were available, all dates were set to the 15th of the month.

4.2. Whole-genome sequence analyses

Whole-genome sequences were analyzed through a variant-calling pipeline developed in house. Trimmomatic v0.39 (Bolger et al., 2014) was used to i) remove the Illumina adapters allowing for 2 mismatches, ii) scan the reads with a 5 bp sliding window approach and trim when the median quality per base drops below 20, and iii) discard reads shorter than 20 bp. For paired-end data, SeqPrep v1.3.1 (https://github.com/jstjohn/SeqPrep) was used to identify and merge reads with an overlap of at least 15 bp. The processed reads were aligned to an inferred ancestor of the MTBC (Comas et al., 2010) using BWA mem v0.7.17 (Li and Durbin, 2009). Duplicate reads were identified and removed using the MarkDuplicates module of Picard v2.26.2 (http://broadinstitute.github.io/picard/). Sequencing reads were taxonomically classified using Kraken (Wood and Salzberg, 2014), and non-*Mtb* mappings were discarded as described previously (Goig et al., 2020). Variant calling was performed using the mutect2 module of GATK v4.2.4.1 (McKenna et al., 2010). Variants were then filtered using the FilterMutectCalls in microbial mode. Supplementary and secondary alignments were excluded (Mariner-Llicer et al., 2024), as well as genomic positions in repetitive regions such as PE, PPE, and PGRS genes or phages (Stucki et al., 2016). Samples with an average sequencing depth lower than 15X or with more than 1 % of contaminating reads from non-tuberculous mycobacteria were excluded from downstream analysis. Lineages were identified based on SNPs as described in Coll et al. (2014).

4.3. Multiple sequence alignments

The VCF of all positions was used to create a consensus fasta sequence per isolate. Chromosomal positions that were covered by less than 7 reads, as well as unfixed positions (variant frequency between 10 % and 90 %), were treated as missing data. An alignment of polymorphic positions was generated for all sequences per lineage per location, by concatenating all high-quality SNPs, excluding sites that had more than 10 % of missing data, as well as drug-resistance-related sites and repetitive regions.

4.4. Clustering rates and terminal branch lengths

A matrix of pairwise TN93 distances between any given two strains was inferred using the variable position alignment. Strains were clustered using the R package cluster (Maechler et al., 2024) with the unweighted pair group average method. A threshold of five SNPs, on average, was used as a cutoff for clustering and clustering rates were calculated as the proportion of clustered strains.

For the calculation of terminal branch lengths, the variable position alignments were augmented with a count of invariant A, C, G, and T nucleotides (Leaché et al., 2015) and used to infer maximum likelihood phylogenies using IQ-TREE (Nguyen et al., 2014). A general time-reversible (GTR) model of sequence evolution was used and the phylogenies were rooted on *Mycobacterium canettii*. The terminal branch length distributions were extracted from the resulting phylogenies.

4.5. Multi-type birth-death model

We fit a multi-type birth-death model to the sequence alignments (Kühnert et al., 2016) (Fig. 1a), where non-infectious and infectious individuals are treated as two different host types. Under this model, only infectious individuals can transmit (occurring at rate λ_M), resulting in a new individual in the 'infected non-infectious' compartment. Individuals in this compartment cannot transmit, but migrate to the infectious compartment at a constant rate σ (we assume no back migration). Infectious individuals become non-infectious due to recovery, death or removal through sampling, occurring at a constant rate δ_M, implying that they get removed from the system. This compartmental setup effectively implies that all individuals who eventually progress to infectious disease first go through a phase of non-infectiousness. Hence, infected individuals who never become infectious are not part of the system under study. Instead of λ_M, the model was parametrized with $R_{e,M}$ (which equals λ_M/δ_M) as more prior knowledge is available for this parameter (Loiseau et al., 2023; Zwyer et al., 2023; Windels et al., 2024). Infected individuals are sampled with sampling proportion p, which was set to zero before the onset of sampling and set to a fixed, non-zero value afterwards. This value was calculated per sampling location as the total number of sequences divided by the number of cases reported during the sampling period and multiplied by 0.25 to reflect 75 % underreporting of TB cases (this underreporting level was varied in the sensitivity analyses; see below). For L6, we took a culture bias into account by assuming that the efficiency of culture growth for L6 is two third of that for L4 (Sanoussi et al., 2017). Upon sampling an infectious patient, the patient was assumed to be removed from the infectious pool with probability r (Gavryushkina et al., 2014). We further assumed a strict molecular clock and a general time-reversible nucleotide substitution model with four gamma rate categories to account for site-to-site rate heterogeneity (GTR+Γ_4). All parameters and their prior distributions are listed in Table S1.

4.6. Single-type birth-death model

The single-type birth-death model (Stadler, 2010) (Fig. 1b) represents a constrained version of the multi-type birth-death model, assuming that individuals instantaneously become infectious upon

infection. Hence, all infected individuals are infectious and transmit at a constant rate λ_S. Individuals become non-infectious at a constant rate δ_S through recovery, death or removal through sampling (with removal probability r). Instead of λ_S, the model was parametrized with $R_{e,S}$, which equals λ_S/δ_S. All other elements of the model, including the sampling, clock, and nucleotide substitution models, are the same as in the multi-type birth-death model. All parameters and their prior distributions are listed in Table S5.

4.7. Phylodynamic inference

We performed phylodynamic inference using the bdmm package (Kühnert et al., 2016) v1.0.3 (https://github.com/tgvaughan/bdmm/releases/tag/v1.0.3-unofficial), feast package v8.3.1 (https://github.com/tgvaughan/feast/releases/tag/v8.3.1), and skylinetools package v0.2.0 (https://github.com/laduplessis/skylinetools/releases/tag/0.2.0) in BEAST v2.6.6 (Bouckaert et al., 2014; Bouckaert et al., 2019). Data from each lineage in each location were analyzed independently. Variable position alignments were augmented with a count of invariant A, C, G, and T nucleotides (Leaché et al., 2015). Alignments containing more than 400 sequences were randomly downsampled to 400 sequences for computational feasibility, and sampling proportions were adjusted accordingly. For each analysis, three independent Markov Chain Monte Carlo chains were run, with states sampled every 1000 steps and trees sampled every 10,000 steps. Convergence was assessed with Tracer (Rambaut et al., 2018), confirming that the effective sample size (ESS) was at least 200 for the parameters of interest. 10 % of each chain was discarded as burn-in, and the remaining samples across the three chains were pooled and downsampled by a factor 10,000 using LogCombiner (Bouckaert et al., 2019), resulting in at least 49,000,000 iterations in combined chains. All phylodynamic inference steps were implemented in a Snakemake workflow (Mölder et al., 2021). Posterior distributions on derived parameters were calculated using the appropriate arithmetic expressions (Table 2). The model assumes that the time until onset of infectiousness is exponentially distributed, with the mean time being the parameter of that exponential distribution. In Fig. 2b, Fig. S3b, Fig. S4b, Fig. S5a, and Fig. S6b we provide the posterior distribution for this mean time until onset of infectiousness, $1/\sigma$. In Fig. 2c, Fig. S3c, Fig. S4c, Fig. S5b, and Fig. S6c we show the exponential distribution for the most likely mean time parameter, i.e. the distribution of all times until infectiousness onset in the population.

4.8. Sensitivity analyses

The robustness of the phylodynamic inference to sampling assumptions was assessed by assuming lower levels of underreporting of TB cases (50 % and 0 %).

To investigate the effect of clock rate assumptions, we focused on L1 and L2 in Tanzania, two lineages with largely different estimates for the non-infectious period (Fig. 2b). We examined how large the difference in clock rates would need to be, if it were invoked as the only factor underlying the difference in branch lengths. To this end, we fixed the clock rate for each lineage to different values, chosen within the range of previously reported values (Menardo et al., 2019), and checked which combination resulted in similar non-infectious period estimates.

When the evolutionary history of the sampled isolates is characterized by multiple independent introduction events in the sampling locations, the early parts of the phylogenetic trees might not result from the same population dynamic process as the more recent parts. This can lead to a model misspecification and might bias the epidemiological parameter estimates. To eliminate such effects, we repeated the multi-type birth-death analyses, allowing for a change in σ and δ_M at a point in time set to 30 years before the most recent sample. For computational feasibility, $R_{e,M}$ was set to 1 for both time intervals.

The Lognormal(0,1) prior distribution on σ in the main multi-type birth-death analyses puts a very small weight on short non-infectious periods. To allow for a non-infectious period close to zero (which would approximately correspond to a single-type birth-death model), we reparametrized the model with $1/\sigma$ and tested the effect of an Exp(1) prior (in other words, a prior with high support for low values) on this parameter.

4.9. Model selection

To perform model selection on the single-type and multi-type birth-death model, we allowed BEAST to select between two models: 1) a model where both σ and the R_e between compartments ($R_{e,M}$) is zero, while the R_e within the infectious compartment ($R_{e,S}$) is non-zero (corresponding to the single-type birth-death model), and 2) a model where $R_{e,S}$ is zero, while σ and $R_{e,M}$ are non-zero (corresponding to the multi-type birth-death model). To this end, spike and slab priors were set on σ, $R_{e,S}$ and $R_{e,M}$, by defining them as ModelSelectionParameter (feast package) and putting the same priors as before on the non-zero values of σ, $R_{e,S}$ and $R_{e,M}$. An equal prior weight was put on both models.

CRediT authorship contribution statement

Windels Etthel M: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Valenzuela Agüí Cecilia:** Writing – review & editing, Software, Methodology. **de Jong Bouke C:** Writing – review & editing, Resources, Conceptualization. **Meehan Conor J:** Writing – review & editing, Data curation. **Loiseau Chloé:** Writing – review & editing, Software. **Goig Galo A:** Writing – review & editing, Software. **Zwyer Michaela:** Writing – review & editing, Software. **Borrell Sonia:** Writing – review & editing. **Brites Daniela:** Writing – review & editing, Supervision, Software, Data curation, Conceptualization. **Gagneux Sebastien:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Stadler Tanja:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare no conflicts of interest.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement no. 101001077 (to E.M.W. and T.S.) and ETH Zürich (to E.M.W. and T.S.). This work was further supported by the Swiss National Science Foundation (grants CRSII5-213514 and 320030-227432 to S.G.) and the ERC (883582 to S.G.). Calculations were performed on the Euler cluster at ETH Zürich and at sciCORE (https://scicore.unibas.ch/) scientific computing core facility at the University of Basel. We would like to thank Timothy G. Vaughan for help with the phylodynamic analyses, Louis du Plessis for insightful discussions, and Christoph Stritt and Antoine Zwaans for valuable feedback on the manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.epidem.2025.100821.

Data availability

A data availability statement is provided in the main text.

References

Asante-Poku, A., Otchere, I.D., Osei-Wusu, S., Sarpong, E., Baddoo, A., Forson, A., Laryea, C., Borrell, S., Bonsu, F., Hattendorf, J., et al., 2016. Molecular epidemiology of *Mycobacterium africanum* in Ghana. BMC Infect. Dis. 16, 385.

Asare, P., Asante-Poku, A., Prah, D.A., Borrell, S., Osei-Wusu, S., Otchere, I.D., Forson, A., Adjapong, G., Koram, K.A., Gagneux, S., et al., 2018. Reduced transmission of *Mycobacterium africanum* compared to *Mycobacterium tuberculosis* in urban West Africa. Int. J. Infect. Dis. 73, 30–42.

Behr, M.A., Edelstein, P.H., Ramakrishnan, L., 2018. Revisiting the timetable of tuberculosis. BMJ 362, k2738.

Behruznia, M., Marin, M., Farhat, M., Thomas, J.C., Domingo-Sananes, R., Meehan, C.J., 2024. The *Mycobacterium tuberculosis* complex pangenome is small and driven by sub-lineage-specific regions of difference. bioRxiv. ⟨https://www.biorxiv.org/content/10.1101/2024.03.12.584580v2⟩.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Borgdorff, M.W., Sebek, M., Geskus, R.B., Kremer, K., Kalisvaart, N., van Soolingen, D., 2011. The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. Int J. Epidemiol. 40, 964–970.

Bottai, D., Frigui, W., Sayes, F., Di Luca, M., Spadoni, D., Pawlik, A., Zoppo, M., Orgeur, M., Khanna, V., Hardy, D., et al., 2020. TbD1 deletion as a driver of the evolutionary success of modern epidemic *Mycobacterium tuberculosis* lineages. Nat. Commun. 11, 684.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10, e1003537.

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., et al., 2019. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15, e1006650.

Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., et al., 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc. Natl. Acad. Sci. 99, 3684–3689.

Cadena, A.M., Fortune, S.M., Flynn, J.L., 2017. Heterogeneity in tuberculosis. Nat. Rev. Immunol. 17, 691–702.

Chen, Y.Y., Chang, J.R., Huang, W.F., Hsu, S.C., Kuo, S.C., Sun, J.R., Dou, H.Y., 2014. The pattern of cytokine production *in vitro* induced by ancient and modern beijing *Mycobacterium tuberculosis* strains. PLoS One 9, e94296.

Colangeli, R., Gupta, A., Vinhas, S.A., Chippada Venkata, U.D., Kim, S., Grady, C., Jones-López, E.C., Soteropoulos, P., Palaci, M., Marques-Rodrigues, P., et al., 2020. *Mycobacterium tuberculosis* progresses through two phases of latent infection in humans. Nat. Commun. 11, 4870.

Coll, F., McNerney, R., Guerra-Assunção, J.A., Glynn, J.R., Perdigão, J., Viveiros, M., Portugal, I., Pain, A., Martin, N., Clark, T.G., 2014. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat. Commun. 5, 4812.

Comas, I., Chakravartti, J., Small, P.M., Galagan, J., Niemann, S., Kremer, K., Ernst, J.D., Gagneux, S., 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat. Genet 42, 498–503.

Coscolla, M., Gagneux, S., 2010. Does *M. tuberculosis* genomic diversity explain disease diversity? Drug Discov. Today Dis. Mech. 7, e43–e59.

Coscolla, M., Gagneux, S., 2014. Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol. 26, 431–444.

Drain, P.K., Bajema, K.L., Dowdy, D., Dheda, K., Naidoo, K., Schumacher, S.G., Ma, S., Meermeier, E., Lewinsohn, D.M., Sherman, D.R., 2018. Incipient and subclinical tuberculosis: a clinical review of early stages and progression of infection. Clin. Microbiol Rev. 31 e00021-18.

Emery, J.C., Richards, A.S., Dale, K.D., McQuaid, C.F., White, R.G., Denholm, J.T., Houben, R.M.G.J., 2021. Self-clearance of *Mycobacterium tuberculosis* infection: implications for lifetime risk and population at-risk of tuberculosis disease. Proc. R. Soc. B 288, 20201635.

Ford, C.B., Lin, P.L., Chase, M.R., Shah, R.R., Iartchouk, O., Galagan, J., Mohaideen, N., Ioerger, T.R., Sacchettini, J.C., Lipsitch, M., et al., 2011. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. Nat. Genet 43, 482–488.

Frascella, B., Richards, A.S., Sossen, B., Emery, J.C., Odone, A., Law, I., Onozaki, I., Esmail, H., Houben, R.M.G.J., 2021. Subclinical tuberculosis disease - a review and analysis of prevalence surveys to inform definitions, burden, associations, and screening methodology. Clin. Infect. Dis. 73, e830–e841.

Freschi, L., Vargas, R., Husain, A., Kamal, S.M.M., Skrahina, A., Tahseen, S., Ismail, N., Barbova, A., Niemann, S., Cirillo, D.M., et al., 2021. Population structure, biogeography and transmissibility of *Mycobacterium tuberculosis*. Nat. Commun. 12, 6099.

Gagneux, S., 2018. Ecology and evolution of *Mycobacterium tuberculosis*. Nat. Rev. Microbiol 16, 202–213.

Gavryushkina, A., Welch, D., Stadler, T., Drummond, A.J., 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Comput. Biol. 10, e1003919.

Gehre, F., Oko, F., Ofori-Anyinam, B., Meehan, J., Windels, E.M., Joof, K., Faal, T., Mendy, F., Jobarteh, T., Gitteh, E., et al., 2024. Assessing the impact of enhanced-case-finding on tuberculosis incidence and transmission in The Gambia using epidemiological and phylodynamic approaches. bioRxiv. ⟨https://www.medrxiv.org/content/10.1101/2024.05.17.24307536v1⟩.

Glynn, J.R., Alghamdi, S., Mallard, K., Mcnerney, R., Ndlovu, R., Munthali, L., Houben, R.M., Fine, P.E.M., French, N., Crampin, A.C., 2010. Changes in *Mycobacterium tuberculosis* genotype families over 20 years in a population-based study in northern Malawi. PLoS One 5, e12259.

Goig, G.A., Blanco, S., Garcia-Basteiro, A.L., Comas, I., 2020. Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. BMC Biol. 18, 24.

Gröschel, M.I., Pérez-Llanos, F.J., Diel, R., Vargas, R., Escuyer, V., Musser, K., Trieu, L., Meissner, J.S., Knorr, J., Klinkenberg, D., et al., 2024. Differential rates of *Mycobacterium tuberculosis* transmission associate with host-pathogen sympatry. Nat. Microbiol 9, 2113–2127.

Guerra-Assunção, J., Crampin, A., Houben, R., Mzembe, T., Mallard, K., Coll, F., Khan, P., Banda, L., Chiwaya, A., Pereira, R., et al., 2015. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. Elife 4, e05166.

Guyeux, C., Senelle, G., Le Meur, A., Supply, P., Gaudin, C., Phelan, J.E., Clark, T.G., Rigouts, L., de Jong, B., Sola, C., et al., 2024. Newly identified *Mycobacterium africanum* lineage 10, Central Africa. Emerg. Infect. Dis. 30, 560–563.

Hiza, H., Zwyer, M., Hella, J., Arbués, A., Sasamalo, M., Borrell, S., Xu, Z.M., Ross, A., Brites, D., Fellay, J., et al., 2024. Bacterial diversity dominates variable macrophage responses of tuberculosis patients in Tanzania. Sci. Rep. 14, 9287.

Holt, K.E., McAdam, P., Thai, P.V.K., Thuong, N.T.T., Ha, D.T.M., Lan, N.N., Lan, N.H., Nhu, N.T.Q., Hai, H.T., Ha, V.T.N., et al., 2018. Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat. Genet 50, 849–856.

Horton, K.C., Richards, A.S., Emery, J.C., Esmail, H., Houben, R.M.G.J., 2023. Reevaluating progression and pathways following *Mycobacterium tuberculosis* infection within the spectrum of tuberculosis. Proc. Natl. Acad. Sci. 120, e2221186120.

de Jong, B.C., Adetifa, I., Walther, B., Hill, P.C., Antonio, M., Ota, M., Adegbola, R.A., 2010. Differences between TB cases infected with *M. africanum*, West-African type 2, relative to Euro-American *M. tuberculosis*- an update. FEMS Immunol. Med Microbiol 58, 102–105.

de Jong, B.C., Hill, P.C., Aiken, A., Awine, T., Antonio, M., Adetifa, I.M., Jackson-Sillah, D.J., Fox, A., DeRiemer, K., Gagneux, S., et al., 2008. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in the Gambia. J. Infect. Dis. 198, 1037–1043.

Kendall, E.A., Shrestha, S., Dowdy, D.W., 2021. The epidemiological importance of subclinical tuberculosis: a critical reappraisal. Am. J. Respir. Crit. Care Med 203, 168–174.

Ku, C.C., MacPherson, P., Khundi, M.E., Nzawa Soko, R.H., Feasey, H.R.A., Nliwasa, M., Horton, K.C., Corbett, E.L., Dodd, P.J., 2021. Durations of asymptomatic, symptomatic, and care-seeking phases of tuberculosis disease with a Bayesian analysis of prevalence survey and notification data. BMC Med 19, 298.

Kühnert, D., Stadler, T., Vaughan, T.G., Drummond, A.J., 2016. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. Mol. Biol. Evol. 33, 2102–2116.

Lau, A., Lin, C., Barrie, J., Winter, C., Armstrong, G., Egedahl, M.Lou, Doroshenko, A., Heffernan, C., Asadi, L., Fisher, D., et al., 2022. The radiographic and mycobacteriologic correlates of subclinical pulmonary TB in Canada: a retrospective cohort study. Chest 162, 309–320.

Leaché, A.D., Banbury, B.L., Felsenstein, J., De Oca, A.N.M., Stamatakis, A., 2015. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. Syst. Biol. 64, 1032–1047.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Lillebaek, T., Norman, A., Rasmussen, E.M., Marvig, R.L., Folkvardsen, D.B., Andersen, Å.B., Jelsbak, L., 2016. Substantial molecular evolution and mutation rates in prolonged latent *Mycobacterium tuberculosis* infection in humans. Int. J. Med. Microbiol. 306, 580–585.

Loiseau, C., Windels, E.M., Gygli, S.M., Jugheli, L., Maghradze, N., Brites, D., Ross, A., Goig, G., Reinhard, M., Borrell, S., et al., 2023. The relative transmission fitness of multidrug-resistant *Mycobacterium tuberculosis* in a drug resistance hotspot. Nat. Commun. 14, 1988.

Long, R., Croxen, M., Lee, R., Doroshenko, A., Lau, A., Asadi, L., Heffernan, C., Paulsen, C., Egedahl, M.Lou, Lloyd, C., et al., 2024. The association between phylogenetic lineage and the subclinical phenotype of pulmonary tuberculosis: a retrospective 2-cohort study. J. Infect. 88, 123–131.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2024. Cluster: cluster analysis basics and extensions. R. Package Version 2 (1), 8.

Mariner-Llicer, C., Goig, G.A., Torres-Puente, M., Vashakidze, S., Villamayor, L.M., Saavedra-Cervera, B., Mambuque, E., Khurtsilava, I., Avaliani, Z., Rosenthal, A., et al., 2024. Genetic diversity within diagnostic sputum samples is mirrored in the culture of *Mycobacterium tuberculosis*. bioRxiv. ⟨https://www.biorxiv.org/content/10.1101/2024.01.30.577772v1.abstract⟩.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al., 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297–1303.

Menardo, F., 2022. Understanding drivers of phylogenetic clustering and terminal branch lengths distribution in epidemics of *Mycobacterium tuberculosis*. Elife 11, e76780.

Menardo, F., Duchêne, S., Brites, D., Gagneux, S., 2019. The molecular clock of *Mycobacterium tuberculosis*. PLoS Pathog. 15, e1008067.

Menzies, N.A., Swartwood, N., Testa, C., Malyuta, Y., Hill, A.N., Marks, S.M., Cohen, T., Salomon, J.A., 2021. Time since infection and risks of future disease for individuals with *Mycobacterium tuberculosis* infection in the United States. Epidemiology 32, 70–78.

Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al., 2021. Sustainable data analysis with Snakemake. F1000Res 10, 33.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.

Orgeur, M., Sous, C., Madacki, J., Brosch, R., 2024. Evolution and emergence of *Mycobacterium tuberculosis*. FEMS Microbiol Rev. 48.

Peters, J.S., Ismail, N., Dippenaar, A., Ma, S., Sherman, D.R., Warren, R.M., Kana, B.D., 2020. Genetic diversity in *Mycobacterium tuberculosis* clinical isolates and resulting outcomes of tuberculosis infection and disease. Annu Rev. Genet 54, 511–537.

Portevin, D., Gagneux, S., Comas, I., Young, D., 2011. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. PLoS Pathog. 7, e1001307.

Rambaut, A., Drummond, A.J., Xie, D., Baele, G., Suchard, M.A., 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67, 901–904.

Reiling, N., Homolka, S., Walter, K., Brandenburg, J., Niwinski, L., Ernst, M., Herzmann, C., Lange, C., Diel, R., Ehlers, S., et al., 2013. Clade-specific virulence patterns of *Mycobacterium tuberculosis* complex strains in human primary macrophages and aerogenically infected mice. mBio 4 e00250-13.

Sanoussi, C.N., Affolabi, D., Rigouts, L., Anagonou, S., de Jong, B., 2017. Genotypic characterization directly applied to sputum improves the detection of *Mycobacterium africanum* West African 1, under-represented in positive cultures. PLoS Negl. Trop. Dis. 11, e0005900.

Sloot, R., Van Der Loeff, M.F.S., Kouw, P.M., Borgdorff, M.W., 2014. Risk of tuberculosis after recent exposure: a 10-year follow-up study of contacts in Amsterdam. Am. J. Respir. Crit. Care Med 190, 1044–1052.

Sobkowiak, B., Banda, L., Mzembe, T., Crampin, A.C., Glynn, J.R., Clark, T.G., 2020. Bayesian reconstruction of *Mycobacterium tuberculosis* transmission networks in a high incidence area over two decades in Malawi reveals associated risk factors and genomic variants. Micro Genom. 6, e000361.

Stadler, T., 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J. Theor. Biol. 261, 58–66.

Stadler, T., 2010. Sampling-through-time in birth-death trees. J. Theor. Biol. 267, 396–404.

Stanley, S., Spaulding, C.N., Liu, Q., Chase, M.R., Ha, D.T.M., Thai, P.V.K., Lan, N.H., Thu, D.D.A., Quang, N., Le, Brown, J., et al., 2024. Identification of bacterial determinants of tuberculosis infection and treatment outcomes: a phenogenomic analysis of clinical strains. Lancet Microbe 5, e570–e580.

Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., Fenner, L., Rutaihwa, L., Borrell, S., Luo, T., et al., 2016. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. Nat. Genet 48, 1535–1543.

Thwaites, G., Caws, M., Chau, T.T.H., D'Sa, A., Lan, N.T.N., Huyen, M.N.T., Gagneux, S., Anh, P.T.H., Dau, Q.T., Torok, E., et al., 2008. Relationship between *Mycobacterium tuberculosis* genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. J. Clin. Microbiol 46, 1363–1368.

Tientcheu, L.D., Koch, A., Ndengane, M., Andoseh, G., Kampmann, B., Wilkinson, R.J., 2017. Immunological consequences of strain variation within the *Mycobacterium tuberculosis* complex. Eur. J. Immunol. 47, 432–445.

Windels, E.M., Wampande, E.M., Joloba, M.L., Boom, W.H., Goig, G.A., Cox, H., Hella, J., Borrell, S., Gagneux, S., Brites, D., et al., 2024. HIV co-infection is associated with reduced *Mycobacterium tuberculosis* transmissibility in sub-Saharan Africa. PLoS Pathog. 20, e1011675.

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46.

World Bank, 2024. Incidence of tuberculosis (per 100,000 people) - The Gambia. World Development Indicators. ⟨https://data.worldbank.org/indicator/SH.TBS.INCD?locations=GM⟩.

Xu, Y., Cancino-Munoz, I., Torres-Puente, M., Villamayor, L.M., Borrás, R., Borrás-Máñez, M., Bosque, M., Camarena, J.J., Colomer-Roig, E., Colomina, J., et al., 2019. High-resolution mapping of tuberculosis transmission: whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. PLoS Med. 16, e1002961.

Yang, C., Luo, T., Sun, Guomei, Qiao, K., Sun, Gang, Deriemer, K., Mei, J., Gao, Q., 2012. *Mycobacterium tuberculosis* Beijing strains favor transmission but not drug resistance in China. Clin. Infect. Dis. 55, 1179–1187.

Zwyer, M., Rutaihwa, L.K., Windels, E., Hella, J., Menardo, F., Sasamalo, M., Sommer, G., Schmülling, L., Borrell, S., Reinhard, M., et al., 2023. Back-to-Africa introductions of *Mycobacterium tuberculosis* as the main cause of tuberculosis in Dar es Salaam, Tanzania. PLoS Pathog. 19, e1010893.