Multiple Introductions of the *Mycobacterium tuberculosis* Lineage 2–Beijing into Africa over centuries

Liliana K. Rutaihwa[1, 2, 3], Fabrizio Menardo[1, 2], David Stucki[1, 2], Sebastian M. Gygli[1, 2], Serej D Ley[1, 2, 4, 5], Bijaya Malla[1, 2, 6], Julia Feldmann[1, 2], Sonia Borrell[1, 2], Christian Beisel[7], Kerren Middelkoop[8, 9], E Jane Carter[10, 11], Lameck Diero[11], Marie Ballif[12] , Levan Jugheli[1, 2], Klaus Reither[1, 2], Lukas Fenner[1, 2, 12], Daniela Brites[1, 2#*] and Sebastien Gagneux[1, 2#*]

[1] Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Basel, Switzerland

[3] Ifakara Health Institute, Bagamoyo, Tanzania

[4] Papua New Guinea Institute of Medical Research, Goroka, Papua New Guinea

[5] DST-NRF Centre of Excellence for Biomedical Tuberculosis Research; South African Medical Research Council Centre for Tuberculosis Research; Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town.

[6] Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland.

[7] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

[8] Desmond Tutu HIV Centre, Institute of Infectious Disease and Molecular Medicine, Cape Town South-Africa

[9] Department of Medicine, University of Cape Town, Cape Town South-Africa

[10] Division of Pulmonary and Critical Care Medicine, Warren Alpert School of Medicine at Brown University, Providence, Rhode Island, USA

[11] Moi University School of Medicine, Eldoret, Kenya

[12] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

[#] Equal contribution

*** Correspondence:** Sebastien Gagneux, sebastien.gagneux@swisstph.ch; Daniela Brites, d.brites@swisstph.ch

Keywords: tuberculosis, introduction, genetic diversity, migration, whole genome sequencing, drug resistance

Abstract

The Lineage 2–Beijing (L2–Beijing) sub-lineage of *Mycobacterium tuberculosis* has received much attention due to its high virulence, fast disease progression, and association with antibiotic resistance. Despite several reports of the recent emergence of L2–Beijing in Africa, no study has investigated the evolutionary history of this sub-lineage on the continent. In this study, we used whole genome sequences of 817 L2 clinical strains from 14 geographical regions globally distributed to investigate the origins and onward spread of this lineage in Africa. Our results reveal multiple introductions of L2–Beijing into Africa linked to independent bacterial populations from East- and Southeast Asia. Bayesian analyses further indicate that these introductions occurred during the past 300 years, with most of these events pre-dating the antibiotic era. Hence, the success of L2–Beijing in Africa is most likely due to its hypervirulence and high transmissibility rather than drug resistance.

1 Introduction

Tuberculosis (TB) is mainly caused by a group of closely related bacteria referred to as the *Mycobacterium tuberculosis* Complex (MTBC). The MTBC comprises seven phylogenetic lineages adapted to humans and several lineages adapted to different wild and domestic animal species (Gagneux, 2018). The human-adapted lineages of the MTBC show a distinct geographic distribution, with some "generalist" lineages such as Lineage (L)2 and L4 occurring all around the world and others being geographically restricted "specialist" that include L5, L6 and L7 (Coscolla and Gagneux, 2014; Stucki *et al.*, 2016). Africa is the only continent which is home to all seven human-adapted lineages, including the three "specialist" lineages exclusively found on the continent. Current evidence suggests that the MTBC overall originated in Africa (Gagneux, 2018) and subsequently spread around the globe following human migratory events (Wirth *et al.*, 2008; Comas *et al.*, 2013). The broad distribution of some of the "generalist" lineages and their presence in Africa has been attributed to past exploration, trade and conquest. For instance, an important part of the TB epidemics in sub-Saharan Africa is driven by the generalist Latin–American–Mediterranean (LAM) sublineage of L4, which is postulated to have been introduced to the continent post-European contact (Stucki *et al.*, 2016).

Among the different human-adapted MTBC lineages, the L2–Beijing sublineage has been of particular interest (Merker *et al.*, 2015). L2–Beijing has expanded and emerged worldwide from East Asia; its most likely geographical origin (Luo *et al.*, 2015; Merker *et al.*, 2015). In some parts of the world, the recent emergence of L2–Beijing has been linked to increased transmission (Yang *et al.*, 2012; Holt *et al.*, 2018), high prevalence of multidrug–resistant TB (MDR–TB) (Borrell and Gagneux, 2009), and to social and political instability, resulting into displacement of people and poor health systems (Eldholm *et al.*, 2016). Increasingly, L2–Beijing is also being reported in Africa (Bifani *et al.*, 2002; Affolabi *et al.*, 2009; Gehre *et al.*, 2016; Mbugi *et al.*, 2016), and evidence suggests that L2–Beijing in African regions is becoming more prevalent over time (Cowley *et al.*, 2008; van der Spuy *et al.*, 2009; Glynn *et al.*, 2010). Some authors have hypothesized that the introduction of L2–Beijing into South Africa resulted from the importation of slaves from Southeast Asia during the 17th and 18th centuries and/or the Chinese labor forces arriving in the 1900s (van Helden *et al.*, 2002). Alternatively, in West Africa, the presence of L2–Beijing was proposed to reflect more recent immigration from Asia (Affolabi *et al.*, 2009; Gehre *et al.*, 2016). To a certain extent, the recent expansion of L2–Beijing in parts of Africa has been associated with drug resistance (Githui *et al.*, 2004; Klopper *et al.*, 2013) and higher transmissibility (Guerra-Assunção *et al.*, 2015). In addition, a study in the Gambia showed a faster progression from latent infection to active TB disease in patient house-hold contacts exposed to L2–Beijing (de Jong *et al.*, 2008).

Whilst L2–Beijing seems to be expanding in several regions of Africa, no study has formally investigated the evolutionary history of L2–Beijing on the continent. In this study, we used whole genome sequencing data from a global collection of L2 clinical strains to determine the most likely geographical origin of L2–Beijing in Africa and its spread across the continent.

2 Results

2.1 Phylogenetic inference of L2 strains

We analyzed a total of 817 L2 genomes originating from 14 geographical regions including Eastern and Southern Africa (Figure S1 and Table S1). We focused on seven geographical regions that had more than 20 genomes each, and assigned the remainder to "Other", including two genomes from Western Africa (Figure 1A). The resulting phylogeny of L2 was divided into two main sublineages: the L2–proto-Beijing and L2–Beijing, supporting previous results (Luo *et al.*, 2015; Shitikov *et al.*, 2017). The L2–proto-Beijing was the most basal L2 sublineage and was restricted to East- and Southeast Asia. L2–Beijing, particularly the "modern" (also known as "typical") sublineage, was geographically widely distributed and included strains from Africa. We further characterized L2–Beijing using the recently described unified classification scheme for L2 (Shitikov *et al.*, 2017).

2.2 The population structure of L2–Beijing in Eastern and Southern Africa

Our findings showed the population of African L2–Beijing to be heterogeneous (Figure 1B, Figure 2 and Table S2). Most of the African L2–Beijing strains were classified into several groups within the "modern" sublineage, which included primarily the "Asian-African" sublineages (L2.2.4, L2.2.5 and L2.2.7), consistent with previous findings (Merker *et al.*, 2015). We also identified the "ancient" (atypical) strains among the African L2–Beijing. Given that "ancient" L2–Beijing strains (L2.2.1 – L2.2.3) are generally uncommon (Luo *et al.*, 2015), it is interesting to observe such strains in both African regions. In several instances, African L2–Beijing strains did not fall into any of the previously defined groups (Figure 2). Of the two African regions studied here, East Africa had higher proportion of previously uncharacterized L2–Beijing strains (50/101, 50%).

In summary, our findings show that African regions harbored distinct L2–Beijing populations. This is unlike Eastern Europe and Central Asia, where L2–Beijing is dominated by a few highly similar strains (Casali *et al.*, 2014; Eldholm *et al.*, 2016). Of note, L2–Beijing strains typical of Central Asia and Eastern Europe were completely absent from the African populations (Figure 2).

2.3 Genetic diversity of L2–Beijing strains across geographic regions

The spatial distribution of L2–Beijing sublineages and the prevalence of "ancient" L2–Beijing strains observed in this study and previously (Luo *et al.*, 2015; Merker *et al.*, 2015), suggest that L2–Beijing has expanded worldwide from Asia. This view can further be supported by the measures of genetic diversity of L2–Beijing in the different geographical regions (Figure 3). As expected, East- and Southeast Asia contained the most genetically diverse L2 populations, which is consistent with previous results (Luo *et al.*, 2015). Conversely, L2 populations in other geographies were less genetically diverse, suggesting recent dissemination of L2 to these regions. Within Africa, Southern Africa showed a higher diversity in L2–Beijing populations compared to Eastern Africa.

The genetic diversity within the African L2–Beijing populations not only reflects the number and variety of source populations but also local patterns of diversification that occurred after their

introduction. Therefore, the higher genetic diversity of the L2–Beijing populations in Southern Africa compared to Eastern Africa likely reflects both aspects.

2.4 Multiple introductions of L2–Beijing from Asia into Africa

Based on our reconstructed phylogeny, African L2–Beijing strains clustered into several unrelated clades indicating multiple introductions into Africa (Figure 1B). We next investigated the most likely geographical origins of those introductions. As anticipated, our ancestral reconstruction estimated East Asia as the most likely origin of all L2 (posterior probability of 96.7%) and L2–Beijing (posterior probability 92.5%). Our data further indicate that L2–Beijing was introduced into Africa from East- and Southeast Asia on multiple occasions independently. Furthermore, we observed both direct introductions from Asia into Africa as well as subsequent dispersal within the continent (Figure S2 and S3).

Using stochastic mapping, we estimated a total of 13 introductions or migration events (M1 – M13) into Africa (Figure 4). Eight of the African L2–Beijing introductions originated from East Asia and five from Southeast Asia. Out of the 13 introductions, three (M3, M10 and M13) were present in both African regions analyzed here, suggesting initial introductions from Asia followed by subsequent spread within Africa. Overall, our analysis inferred more independent introductions into Southern Africa than Eastern Africa, seven (M1, M4, M7-9, M11 and M12) and three (M2, M5 and M6), respectively. Taken together, our data suggest that multiple migration events have shaped the populations of L2–Beijing in Africa.

2.5 Bayesian molecular dating

Different hypotheses have been formulated on the possible timing of the introduction of L2–Beijing into Africa (van Helden *et al.*, 2002). Here we used tip-calibration to date the phylogenetic tree of L2 and estimate the age of its introduction to Africa. For these analyses, we identified 308 strains among the 817 for which the sampling year was known. These strains were sampled during a period of 19 years; 1995 - 2014 (Figure S4), were evenly distributed on the complete phylogenetic tree (Figure S5) and included 40% members of the African L2–Beijing strains (Figure S6). Eleven of the 13 African introductions were represented in this dataset (M1-M3 and M6-M13).

We detected no overlap in the 95% credibility interval of the clock rate estimates of observed and randomized datasets indicating that there was sufficient temporal signal in the dataset to perform inference (see methods, Figure S7). Further, We found that the UCLD clock had the highest marginal likelihood and a Bayes Factor of 27 with the second best fitting model, the strict clock (Table 1), indicating strong evidence in favor of the UCLD clock (Kass and Raftery, 1995).

We performed a phylogenetic analysis with BEAST2 using the UCLD clock, and repeated the tip randomization test (see methods). We found that the 95% credibility interval of the clock rate estimates of observed and randomized datasets did not overlap, confirming that there was a temporal signal in our sequence data (Figure S8). Additionally, under the UCLD model, the coefficient of variation (COV), which is a summary of the branch rates distribution (standard deviation divided by

the mean), gives an indication on the clock-likeness of the data (Drummond *et al.*, 2006). A coefficient of variation of zero indicates that the data fit a strict clock, whilst a greater COV indicates a higher heterogeneity of rates through the phylogeny. We obtained a mean COV of 0.22 (95% credibility interval= 0.1732, 0.2732), indicating a moderate level of rate variation across different branches and thus supporting the results of the path sampling analysis that favored the UCLD model.

2.6 Recent origins of the African L2–Beijing clades

We used the UCLD clock model to infer the clock rate and divergent times of the 308 L2 strains with known sampling dates and estimated a mean substitution rate of 1.34×10^{-7} [95% Highest Posterior Density (HPD), $9.2867 \times 10^{-}8$ - 1.7719×10^{-7}]. These estimates are in agreement with previously reported rates from epidemiological studies (Walker *et al.*, 2013; Eldholm *et al.*, 2015). We estimated the most recent common ancestor (MRCA) of the extant L2–Beijing of the 308 strains to the year 1225 [95% HPD, 900 - 1519] (Figure S9). For each African clade, we estimated the year of introduction using the 0.975 quantile of the HPD of the age of the MRCA as the upper limit (most recent possible year) and the 0.025 quantile of the HPD of the divergence time between the closest non-African L2–Beijing strain (the closest outgroup) and the African clade of interest as lowest limit (most ancient possible year). Our estimates placed the earliest introductions of the African L2–Beijing (M1, M3, M7, and M12) in the 18th and 19th century (Figure 5 and Table S3). Four additional migration events (M6, M9, M10, and M11) were estimated to have occurred between the beginning of the 19th century and the first half of the 20th century. Finally, the three most recent introductions to Africa happened in the second half of the 20th century (M2, M8, and M13). Diversity patterns of the African clades exclusive to Eastern- and Southern Africa could further provide support for the recent introductions of African L2–Beijing. We thus calculated the pairwise SNP distances within the individual introductions to explore the local patterns of diversification associated with regional epidemics after the introductions. Although strains within Southern African introductions were relatively more distantly related, L2–Beijing strains from both African regions were on average 20 to 40 SNPs apart (Figure S10 and S11). The latter thresholds were proposed to correspond to strains involved in transmission clusters of estimated 50 to100 years (Meehan *et al.*, 2018), supporting the relatively recent introductions of L2–Beijing into the African continent.

Overall, these results indicate that the different introductions of L2–Beijing to Africa occurred over a period of 300 years. While the earliest introduction is unlikely to have happened after 1732 - 1874, the most recent is unlikely to have occurred before 1946 - 1980.

2.7 Introductions of L2–Beijing into Africa unrelated to drug resistance

Because of the repeated association of L2–Beijing with antibiotic resistance (Borrell and Gagneux, 2009), the emergence and dissemination of L2–Beijing strains has been attributed to drug resistance. However, our estimated timing of these introductions suggest that African L2–Beijing strains were introduced prior the discovery of TB antibiotics, and thus must have involved drug-susceptible strains (Figure 5). To explore this question further, we assessed the drug resistance profiles of L2–Beijing strains linked to the various introduction events into the two African regions. We found that

all the Eastern African populations contained only drug-susceptible strains and that approximately three-quarters of L2–Beijing strains in the Southern African populations were drug-susceptible, with the remaining being either mono–, multi– or extensively drug-resistant (Figure 6 and Figure S12). Taken together, these results suggest that the emergence of L2–Beijing in Africa, particularly in Eastern Africa, was not driven by drug resistance. Moreover, our data indicate independent acquisition of drug resistance for the resistant strains detected in the Southern African L2–Beijing population (Figure 6), which might partly contribute to the subsequent spread of L2–Beijing in Southern Africa but not in Eastern Africa.

3 Discussion

This study investigated the most likely geographical origin of the L2–Beijing in Africa. In line with previous findings (Luo *et al.*, 2015; Merker *et al.*, 2015), we identified East Asia as the most likely place of origin of L2 and L2–Beijing. Our findings further revealed multiple independent introductions of L2–Beijing into Africa linked to separate populations originating from both East- and Southeast Asia. Some of these introductions were followed by further onward spread of L2–Beijing within African regions. Finally, we demonstrate that most introductions of L2–Beijing on the continent occurred before the age of antibiotics.

L2–Beijing has received much attention given its hypervirulence in infection models (Manca *et al.*, 2001; Ribeiro *et al.*, 2014), faster progression to disease and higher transmission potential in humans (de Jong *et al.*, 2008; Holt *et al.*, 2018), frequent association with drug resistance, and recent emergence in different regions of the world (Bifani *et al.*, 2002; Borrell and Gagneux, 2009; Fenner *et al.*, 2013). Several studies indicate L2–Beijing originated in Asia and spread from there to the rest of the world (Luo *et al.*, 2015; Merker *et al.*, 2015). Our results support this notion by identifying "Asia" as the most likely geographical origin of both L2 and L2–Beijing based on our ancestral reconstructions and the fact the L2–Beijing populations in Asia are much more diverse than in other regions. In addition, our findings show that L2–Beijing was introduced into Africa multiple times from both East- and Southeast Asia. The presence of L2–Beijing in South Africa has previously been proposed to be due to the importation of slaves from Southeastern Asia by Europeans in the 17th and 18th centuries followed by the import of Chinese labor-forces in the early 1900s (van Helden *et al.*, 2002; Mokrousov *et al.*, 2005). Our Bayesian dating estimates predicted the earliest introductions of L2– Beijing into Africa to have occurred in the 18th and 19th centuries, concurring with these proposed time periods. However, our findings also point to later introductions of L2–Beijing into the continent in the 19th and early 20th centuries. The timings of the latest three introductions in the second half of the 20th century coincide with the decolonization and post-colonial period in Africa when investments into infrastructure and other projects by Chinese enterprises substantially increased (Yuan, 2006; Rice, 2011). These activities also brought many Chinese workers to Africa during a time when TB was still very prevalent in China (Murray, 2018). Hence, many of these workers were likely latently infected with L2–Beijing and might have later reactivated (Moreira Pescarini *et al.*, 2017). Overall, our findings suggest that L2–Beijing has emerged in Africa over the last 300 years and that these introductions have occurred sporadically ever since.

The repeated association of L2–Beijing with drug resistance (Borrell and Gagneux, 2009) has led some to propose that drug resistance is another reason why this sublineage might successfully compete against and eventually replace other *M. tuberculosis* genotypes (Parwati, Van Crevel and Van Soolingen, 2010). However, the underlying reasons for the association of L2–Beijing with drug resistance remains unclear (Borrell and Trauner, 2017), and it is also far from universal, with several reports from e.g. China and other regions finding no such association (Hanekom *et al.*, 2007; Yang *et al.*, 2012). Our results show that most introduction events of L2–Beijing into Africa pre-date the antibiotic era, and because of that these introductions were most likely caused by drug-susceptible strains. The notion that the initial emergence of L2–Beijing in Africa was not driven by drug

resistance is further supported by our findings that none of L2–Beijing strains from Eastern Africa strains analyzed here were drug-resistant. Of note, our observations suggest that drug resistance in South Africa was acquired via independent events post initial introductions from Asia. This is in sharp contrast to the situation in Eastern Europe and Central Asia, where L2–Beijing is highly prevalent but dominated by few recently expanded drug-resistant clones, which account for up to 60% of the L2–Beijing populations in some of these countries (Casali *et al.*, 2014; Eldholm *et al.*, 2016). The association of L2–Beijing with drug resistance in these regions were likely favored by the economic and public health crises that followed the collapse of Soviet Union (Luo *et al.*, 2015; Merker *et al.*, 2015).

Based on our finding that the original introductions of L2–Beijing into Africa involved drug-susceptible strains and that the prevalence of drug-resistant L2–Beijing in Africa overall is comparably low (WHO, 2017), we propose that some of the other characteristics of this sub-lineage, in particular its high virulence, high transmissibility and rapid progression from infection to disease, were responsible for the initial competitive success of L2–Beijing in Africa. Given that the MTBC overall originated in Africa (Wirth *et al.*, 2008; Comas *et al.*, 2013), TB epidemics on the continent were caused by many different "native" genotypes prior to foreign contacts (Comas *et al.*, 2015). The emergence and expansion of "foreign" genotypes including L2–Beijing post-contact demonstrate their ability to successfully compete against the existing genotypes on the continent, irrespective of drug resistance. Following their initial establishment, poor TB treatment programs subsequently selected for drug resistance in L2–Beijing but also in other MTBC lineages, which might have facilitated their further spread in countries such as South Africa (Müller *et al.*, 2013).

This study is limited by the fact that we analyzed a globally diverse collection of L2 genomes available in public repositories. Hence, these strains might not be fully representative of the respective geographical regions. Moreover, our African L2–Beijing dataset came from convenient sampling and comprised L2–Beijing mainly from Eastern and Southern Africa, as whole genome data of L2–Beijing from the other African regions were unavailable at the time of the study. However, the representation of African L2–Beijing in our sample reflects the overall prevalence of this sub-lineage as recently reported for the continent (Mbugi *et al.*, 2016; Chihota *et al.*, 2018). Moreover, although regions outside of Eastern- and Southern Africa were underrepresented, this is unlikely to invalidate our findings regarding the multiple independents of L2–Beijing into Africa, except by underestimating the number of true introductions.

In conclusion, this is the first study to address the geographical origins of L2–Beijing in Africa using whole genome sequencing data. Our findings indicate multiple independent introductions of L2–Beijing epidemics into Africa from East- and Southeast Asia during the last 300 years that were unrelated to drug resistance. The TB epidemics in Africa have remained fairly stable over the last few decades (WHO, 2017). However, Africa's population growth and increasing urbanization (driven by booming economies) are likely to have an impact on the future of TB in this continent, whether directly by e.g. facilitating transmission or indirectly by promoting new risk factors such as diabetes that increase TB susceptibility (Dye and Williams, 2010). It is therefore crucial to follow the TB epidemics in the continent very closely, especially those related to hypervirulent strains such as L2–

Beijing, as these might take particular advantage of this expanding ecological niche (Cowley *et al.*, 2008).

4 Material and Methods

4.1 Identification of Lineage 2 strains and whole-genome sequencing

We obtained whole-genome sequencing data of L2 strains from the two previously largest studies focusing on the evolutionary history and global spread of L2–Beijing strains (Luo *et al.*, 2015; Merker *et al.*, 2015). We then identified additional published genomes as African representatives of L2–Beijing strains from other studies (Guerra-Assunção *et al.*, 2015; Manson *et al.*, 2017). Moreover, we newly sequenced 116 additional L2–Beijing strains using Illumina HiSeq 2000/2500 paired end technology (PRJNA488343). In total, we included 817 L2 genome sequences (Figure S1 and Table S1).

4.2 Whole genome sequence analysis and phylogenetic inference

We used a customized pipeline previously described to map short sequencing reads with BWA 0.6.2 to a reconstructed hypothetical MTBC ancestor used as reference (Comas *et al.*, 2013). SAMtools 0.1.19 was used to call single nucleotide polymorphisms (SNPs), and these SNPs were annotated using ANNOVAR and customized scripts based on the *M. tuberculosis* H37Rv reference annotation (AL123456.2). For downstream analyses, we excluded SNPs in repetitive regions, those annotated in problematic regions such as 'PE/PPE/PGRS' and SNPs in drug-resistance associated genes. Small insertions and deletions were also excluded from the analyses. Only SNPs with minimum coverage of 20x and minimum mapping quality of 30 were kept. All SPNs classified by Samtools as having frequencies of the major non-reference allele lower than 100% (AF1<1) within each genome were considered to be heterogeneous and were treated as ambiguities, and otherwise considered fixed (AF1=1). We concatenated fixed SNPs from the variable positions obtained, which yielded a 33,776 bp alignment. The alignment was then used to infer a maximum likelihood phylogeny using RAxML 8.3.2 with a general time reversible (GTR) model in RAxML and 1,000 rapid bootstrap inferences, followed by a thorough maximum-likelihood search (Stamatakis, 2006).

4.3 Phylogeographic analyses

4.3.1 Reconstruction of ancestral state

To investigate the likely geographic origin of L2–Beijing strains in Africa, we inferred the historical biogeography of L2 using the RASP software (Yu *et al.*, 2015) on a representative subset of 430 genomes due to software's sample limitation. We achieved this subset by performing hierarchical clustering implemented in *pvclust* package in R (Suzuki and Shimodaira, 2006) on a distance matrix and randomly removing clustered genomes. We applied a Bayesian based method in RASP to reconstruct ancestral geographical states on the phylogeny of 430 L2 genomes. We used geographical regions as proxy for origins of the L2 strains and loaded them as distributions. We then ran Bayesian analysis with 5 chains and 500 generations.

4.3.2 Stochastic character mapping

To determine the number of introduction events of L2–Beijing into African regions, we applied stochastic mapping (SIMMAP) on the 817 L2 phylogeny using phytools package 0.6.44 in R (Revell,

2012). Geographical origin of the L2 strains was treated as a discrete trait and modeled onto the phylogeny using ARD model with 100 replicates. This model permits independent region-to-region transfer. We referred to the resulting introductions as migration events "M", considering only those introductions with more than 5 genomes.

4.3.3 Population genetic analyses

4.3.3.1 Nucleotide diversity (pi)

We calculated the mean pair-wise nucleotide diversity per site (Pi) measured by geographic region. We excluded geographic regions represented by less than 20 genomes. Confidence intervals were obtained by bootstrapping through resampling using the sample function in R with replacement and the respective lower and upper confidence levels by calculating 2.5th and 97.5th quartiles.

4.3.3.2 Pairwise SNP distances

We used *dist.dna* function of ape package implemented in R (Paradis, Claude and Strimmer, 2004) to calculate pairwise SNP distances with raw mutation counts and pairwise deletions for gaps. Mean pairwise SNP distance to all strains of the same geographic population was calculated per strain and the distribution of the mean SNP pairwise distance plotted. The mean pairwise SNP distances were assumed not to be normally distributed and we therefore used Wilcoxon rank-sum test to test the differences among geographic regions. Additionally, we calculated pairwise SNP distances within African L2–Beijing populations for migration events with more than 10 genomes each.

4.3.4 Drug Resistance

To distinguish between drug-susceptible and drug-resistant strains, we used genotypic drug resistance molecular markers previously described (Steiner *et al.*, 2014). We categorized strains into: susceptible as having no drug resistance specific mutations; monoresistant as having mutations conferring resistance to a single drug; MDR as having mutations conferring resistance to isoniazid and rifampicin; and extensively drug-resistant (XDR) as having mutations conferring resistance to fluoroquinolones and aminoglycosides in addition to being MDR (Table S4).

4.3.5 Bayesian molecular dating

4.3.5.1 Data preparation and preliminary analysis

To estimate the historical period in which L2–Beijing was introduced to Africa, we performed a set of Bayesian phylogenetic analyses using tip-calibration (Rieux and Khatchikian, 2017). Among the 817 studied L2 strains, we had information on the year of sampling for 308. We performed all further analysis on this subset of 308 strains. We excluded all genomic positions that were invariable in this subset and all positions that were undetermined (missing data or deletions) in more than 25% of the strains, and obtained an alignment of 10,769 polymorphic positions.

In tip dating analysis it is important to test whether the dataset contains strong enough temporal signal (Rieux and Balloux, 2016). To do this, we performed a tip randomization test (Ramsden *et al.*, 2008) as follows. We used BEAST2 v. 2.4.8 (Bouckaert *et al.*, 2014) to run a phylogenetic analysis

with a HKY + GAMMA model (Hasegawa, Kishino and Yano, 1985), a constant population size prior on the tree and a strict molecular clock. Additionally, we used the years in which the strains were sampled to time-calibrate the tree, and we modified the extensible markup language (xml) file to specify the number of invariant sites as indicated by the developers of BEAST2 here: https://groups.google.com/forum/#!topic/beast-users/QfBHMOqImFE (*strict_preliminary.xml*). We ran three independent runs (245 million generations in total), and we used Tracer 1.7 (Rambaut *et al.*, 2018) to identify the burn-in (8 million generations), to assess that the different runs converged, and to estimate the effective sample size (ESS) for all parameters, the posterior and the likelihood (ESS > 110 for all parameters). We then used TipDatingBeast (Rieux and Khatchikian, 2017) to generate 20 replicates of the xml file in which the sampling dates were randomly reassigned to different strains. For each replication, we ran the same BEAST2 analysis as for the original (observed) dataset (one run per replicate, 50 million generations, 10% burn-in). We used TipDatingBeast to parse the log files output of BEAST2 and compare the clock rate estimates for the observed data and the randomized replications. The estimates of the molecular clock rate did not overlap between the observed and the randomized dataset, indicating that there is a clear temporal signal and that we could proceed with further analysis (Figure S5).

4.3.5.2 Model selection

To identify the model that best fits the data, we estimated the marginal likelihood of three different clock models: UCED and UCLD (Drummond *et al.*, 2006). We used the Model selection package of BEAST2 to run a path sampling analysis (Lartillot and Philippe, 2006) following the recommendations of the BEAST2 developers (http://www.beast2.org/path-sampling/). We used the following settings: 100 steps, 4 million generations per step, alpha = 0.3, pre-burn-in = 1 million generations, burn-in for each step= 40% (**PS.xml*). For these analyses, we used proper priors as suggested by (Baele *et al.*, 2012).

4.3.5.3 UCLD analysis

Since the model selection analysis indicated that the UCLD clock was the best fitting model, we repeated the analysis using the UCLD and the same settings used in the path sampling analysis, sampling every 10,000 generations. We ran three independent runs (800 million generations in total), we used Tracer 1.7 (Rambaut *et al.*, 2018) to identify the burn-in (10 million generations), to assess that the different runs converged and to estimate the effective sample size (ESS) for all parameters, the posterior and the likelihood (ESS > 260 for all parameters) (*UCLD_final.xml* and Table S5)

We checked the sensitivity to the priors by running one analysis of 250 million generation sampling from the prior, and compared the parameter estimates with the analysis using the data. We observed the posterior distribution and the prior distribution of all parameters are very distinct (Table S6), indicating that the parameter estimates are influenced by the data and not by the priors (Bromham *et al.*, 2018).

We repeated the tip randomization test with the UCLD model as described above (20 replicates, one run per replicate, 105 million generations per replicate or more, burn-in 10%), and again we found a temporal signal (Figure S8).

To summarize the results, we sampled the trees from the three runs (5% burn-in corresponding to 10 million generations or more, sampling every 25,000 generation). We then summarized the 31,758 sampled trees, created a maximum clade credibility tree using the software TreeAnnotator from the BEAST2 package and used FigTree version 1.4.2 for visualization (Figure S9).

5 Conflict of Interest

We declare no conflict of interest.

6 Author Contributions

LKR, DB, FM, DS, LF and SG planned the study, SDL, BM and JF performed the experiments, LKR, DB, FM, SMG, SDL, BM, CB, SB, KM, MB, LJ, KR and LF contributed strains and prepared the data, LKR, DB, FM and SG analyzed the data, LKR, DB, FM and SG drafted the manuscript. All authors critically reviewed the manuscript.

7 Acknowledgments and funding information

We would like to thank Sebastián Duchêne and Yan Yu for their technical support and Linda-Gail Bekker for contributing strains. All bioinformatics analyses were performed at the scientific computing core facility of the University of Basel, sciCORE (http://scicore.unibas.ch/). This work was supported by the Swiss National Science Foundation (grants 310030_166687 to SG), the European Research Council (309540-EVODRTB to SG) and SystemsX.ch This research was also partially supported (strain collection) by a funding supplement from the National Institutes of Allergy and Infectious Diseases (NIAID) under award numbers U01 AI069924 (IeDEA Southern Africa) and U01 AI069911 (IeDEA East Africa).

8 References

Affolabi, D. *et al.* (2009) 'Possible Outbreak of Streptomycin-Resistant *Mycobacterium tuberculosis* Beijing in Benin', *Emerging Infectious Diseases*, 15(7), pp. 1123–1125. doi: 10.3201/eid1507.080697.

Baele, G. *et al.* (2012) 'Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics', *Molecular Biology and Evolution*. Oxford University Press, 30(2), pp. 239–243. doi: 10.1093/molbev/mss243.

Bifani, P. J. *et al.* (2002) 'Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains', *Trends in Microbiology*, 10(1), pp. 45–52. doi: 10.1016/S0966-842X(01)02277-6.

Borrell, S. and Gagneux, S. (2009) 'Infectiousness, reproductive fitness and evolution of drug-resistant Mycobacterium tuberculosis', *Int J Tuberc Lung Dis*, 13(12), pp. 1456–1466.

Borrell, S. and Trauner, A. (2017) 'Strain Diversity and the Evolution of Antibiotic Resistance', in Gagneux, S. (ed.) *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control*. Cham: Springer International Publishing, pp. 263–279. doi: 10.1007/978-3-319-64371-7_14.

Bouckaert, R. *et al.* (2014) 'BEAST 2: A Software Platform for Bayesian Evolutionary Analysis', *PLoS Comput Biol*, 10(4), p. 1003537. doi: 10.1371/journal.pcbi.1003537.

Bromham, L. *et al.* (2018) 'Bayesian molecular dating: opening up the black box', *Biological Reviews*. Wiley/Blackwell (10.1111), 93(2), pp. 1165–1191. doi: 10.1111/brv.12390.

Casali, N. *et al.* (2014) 'Evolution and transmission of drug-resistant tuberculosis in a Russian population', *Nature Publishing Group*, 46(3), pp. 279–86. doi: 10.1038/ng.2878.

Chihota, V. N. *et al.* (2018) 'Geospatial distribution of Mycobacterium tuberculosis genotypes in Africa', *PLOS ONE*, 13(8), p. e0200632. doi: 10.1371/journal.pone.0200632.

Comas, I. *et al.* (2013) 'Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans', *Nature Publishing Group*, 45(10). doi: 10.1038/ng.2744.

Comas, I. *et al.* (2015) 'Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa', *Current Biology*, 25(24), pp. 3260–3266. doi: 10.1016/j.cub.2015.10.061.

Coscolla, M. and Gagneux, S. (2014) 'Consequences of genomic diversity in Mycobacterium tuberculosis', *Seminars in Immunology*. (Immunity to Mycobacterium tuberculosis), 26(6), pp. 431–444. doi: 10.1016/j.smim.2014.09.012.

Cowley, D. *et al.* (2008) 'Recent and Rapid Emergence of W-Beijing Strains of *Mycobacterium tuberculosis* in Cape Town, South Africa', *Clinical Infectious Diseases*. Oxford University Press, 47(10), pp. 1252–1259. doi: 10.1086/592575.

Drummond, A. J. *et al.* (2006) 'Relaxed phylogenetics and dating with confidence', *PLoS Biol*, 4(5), p. 88. doi: 10.1371/journal.pbio.0040088.

Dye, C. and Williams, B. G. (2010) 'The Population Dynamics and Control of Tuberculosis', *Science*, 328(5980), pp. 856–861. doi: 10.1126/science.1185449.

Eldholm, V. *et al.* (2015) 'Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain', *Nature Communications*, 6(1), p. 7119. doi: 10.1038/ncomms8119.

Eldholm, V. *et al.* (2016) 'Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 113(48), pp. 13881–13886. doi: 10.1073/PNAS.1611283113.

Fenner, L. *et al.* (2013) 'HIV Infection Disrupts the Sympatric Host–Pathogen Relationship in Human Tuberculosis', *PLoS Genet*, 9(3), p. e1003318. doi: 10.1371/journal.pgen.1003318.

Gagneux, S. (2018) 'Ecology and evolution of Mycobacterium tuberculosis', *Nature Publishing Group*, 16. doi: 10.1038/nrmicro.2018.8.

Gehre, F. *et al.* (2016) 'A Mycobacterial Perspective on Tuberculosis in West Africa: Significant Geographical Variation of M. africanum and Other M. tuberculosis Complex Lineages', *PLoS Negl Trop Dis*, 10(3), p. e0004408. doi: 10.1371/journal.pntd.0004408.

Githui, W. A. *et al.* (2004) 'Identification of MDR-TB Beijing/W and other Mycobacterium tuberculosis genotypes in Nairobi, Kenya.', *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 8(3), pp. 352–60.

Glynn, J. R. *et al.* (2010) 'Changes in Mycobacterium tuberculosis Genotype Families Over 20 Years in a Population Based Study in Northern Malawi', *PLoS ONE*, 5(8), p. e12259. doi: 10.1371/journal.pone.0012259.

Guerra-Assunção, J. *et al.* (2015) 'Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area', *Elife*, 4, pp. 285–292. doi: 10.7554/eLife.05166.

Hanekom, M. *et al.* (2007) 'A recently evolved sublineage of the Mycobacterium tuberculosis Beijing strain family is associated with an increased ability to spread and cause disease.', *Journal of clinical microbiology*. American Society for Microbiology, 45(5), pp. 1483–90. doi: 10.1128/JCM.02191-06.

Hasegawa, M., Kishino, H. and Yano, T. (1985) 'Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.', *Journal of molecular evolution*, 22(2), pp. 160–74.

van Helden, P. D. *et al.* (2002) 'Strain families of Mycobacterium tuberculosis', *Trends in Microbiology*, 10(4), pp. 167–168. doi: 10.1016/S0966-842X(02)02317-X.

Holt, K. E. *et al.* (2018) 'Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam', *Nature Genetics*, 50(6), pp. 849–856. doi: 10.1038/s41588-018-0117-9.

de Jong, B. C. *et al.* (2008) 'Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia.', *The Journal of infectious diseases*, 198(7), pp. 1037–43. doi: 10.1086/591504.

Kass, R. E. and Raftery, A. E. (1995) 'Bayes Factors', *Journal of the American Statistical Association*, 90(430), pp. 773–795. doi: 10.1080/01621459.1995.10476572.

Klopper, M. *et al.* (2013) 'Emergence and Spread of Extensively and Totally Drug-Resistant Tuberculosis, South Africa', *Emerging Infectious Diseases*, 19(3), pp. 449–455. doi: 10.3201//EID1903.120246.

Lartillot, N. and Philippe, H. (2006) 'Computing Bayes Factors Using Thermodynamic Integration', *Systematic Biology*. Edited by P. Lewis. Oxford University Press, 55(2), pp. 195–207. doi: 10.1080/10635150500433722.

Luo, T. *et al.* (2015) 'Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese', *PNAS*, p. 201424063. doi: 10.1073/pnas.1424063112.

Manca, C. *et al.* (2001) 'Virulence of a Mycobacterium tuberculosis clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-α/β', *Proceedings of the National Academy of Sciences*, 98(10), pp. 5752–5757.

Manson, A. L. *et al.* (2017) 'Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance', *Nature GeNetics*, 49(3). doi: 10.1038/ng.3767.

Mbugi, E. V. *et al.* (2016) 'Mapping of Mycobacterium tuberculosis Complex Genetic Diversity Profiles in Tanzania and Other African Countries', *PLOS ONE*. Edited by S. Sreevatsan. World Bank, 11(5), p. e0154571. doi: 10.1371/journal.pone.0154571.

Meehan, C. J. *et al.* (2018) 'The relationship between transmission time and clustering methods in 1 Mycobacterium 2', *bioRxiv*. doi: 10.1101/302232.

Merker, M. *et al.* (2015) 'Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage', *Nature Publishing Group*, 47. doi: 10.1038/ng.3195.

Mokrousov, I. *et al.* (2005) 'Origin and primary dispersal of the Mycobacterium tuberculosis Beijing genotype: clues from human phylogeography.', *Genome research*. Cold Spring Harbor Laboratory Press, 15(10), pp. 1357–64. doi: 10.1101/gr.3840605.

Moreira Pescarini, J. *et al.* (2017) 'Migration to middle-income countries and tuberculosis-global policies for global economies', *Globalization and Health*, 13(1), p. 15. doi: 10.1186/s12992-017-0236-6.

Müller, B. *et al.* (2013) 'The heterogeneous evolution of multidrug-resistant Mycobacterium tuberculosis', *Trends in Genetics*, 29(3), pp. 160–169. doi: 10.1016/j.tig.2012.11.005.

Murray, J. F. (2018) 'Tuberculosis in China before, during, and after the Sino-Japanese War', in *Tuberculosis and War*. Karger Publishers, pp. 204–212. doi: 10.1159/000481489.

Paradis, E., Claude, J. and Strimmer, K. (2004) 'APE: Analyses of Phylogenetics and Evolution in R language', *Bioinformatics*. Oxford University Press, 20(2), pp. 289–290. doi: 10.1093/bioinformatics/btg412.

Parwati, I., Van Crevel, R. and Van Soolingen, D. (2010) 'Possible underlying mechanisms for successful emergence of the Mycobacterium tuberculosis Beijing genotype strains'. doi: 10.1016/S1473-3099(09)70330-5.

Rambaut, A. *et al.* (2018) 'Software for Systematics and Evolution Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Systematic Biology*, 0(0), pp. 1–5. doi: 10.1093/sysbio/syy032.

Ramsden, C. *et al.* (2008) 'High Rates of Molecular Evolution in Hantaviruses', *Molecular Biology and Evolution*, 25(7), pp. 1488–1492. doi: 10.1093/molbev/msn093.

Revell, L. J. (2012) 'phytools: an R package for phylogenetic comparative biology (and other things)', *Methods in Ecology and Evolution*. Wiley/Blackwell (10.1111), 3(2), pp. 217–223. doi: 10.1111/j.2041-210X.2011.00169.x.

Ribeiro, S. C. M. *et al.* (2014) 'Mycobacterium tuberculosis strains of the modern sublineage of the Beijing family are more likely to display increased virulence than strains of the ancient sublineage', *J. Clin. Microbiol.*, 52(7), pp. 2615–2624. doi: 10.1128/JCM.00498-14.

Rice, X. (2011) 'China ' s economic invasion of Africa', *Africa*, pp. 1–8.

Rieux, A. and Balloux, F. (2016) 'Inferences from tip-calibrated phylogenies: a review and a practical guide', *Molecular Ecology*, 25(9), pp. 1911–1924. doi: 10.1111/mec.13586.

Rieux, A. and Khatchikian, C. E. (2017) 'tipdatingbeast : an r package to assist the implementation of phylogenetic tip-dating tests using beast', *Molecular Ecology Resources*. Wiley/Blackwell (10.1111), 17(4), pp. 608–613. doi: 10.1111/1755-0998.12603.

Shitikov, E. *et al.* (2017) 'Evolutionary pathway analysis and unified classification of East Asian lineage of Mycobacterium tuberculosis', *Scientific Reports*, 7(1), p. 9227. doi: 10.1038/s41598-017-10018-5.

van der Spuy, G. D. *et al.* (2009) 'Changing Mycobacterium tuberculosis population highlights clade-specific pathogenic characteristics', *Tuberculosis*. Churchill Livingstone, 89(2), pp. 120–125. doi: 10.1016/J.TUBE.2008.09.003.

Stamatakis, A. (2006) 'RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models', *Bioinformatics*, 22(21), pp. 2688–2690. doi: 10.1093/bioinformatics/btl446.

Steiner, A. *et al.* (2014) 'KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes', *BMC Genomics*, 15(1), p. 881. doi: 10.1186/1471-2164-15-881.

Stucki, D. *et al.* (2016) 'Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages', *Nature Genetics*, 48(12). doi: 10.1038/ng.3704.

Suzuki, R. and Shimodaira, H. (2006) 'Pvclust: an R package for assessing the uncertainty in hierarchical clustering', *Bioinformatics*. Oxford University Press, 22(12), pp. 1540–1542. doi: 10.1093/bioinformatics/btl117.

563 Walker, T. M. *et al.* (2013) 'Whole-genome sequencing to delineate Mycobacterium tuberculosis
564 outbreaks: a retrospective observational study.', *The Lancet. Infectious diseases*, 13(2), pp. 137–46.
565 doi: 10.1016/S1473-3099(12)70277-3.

566 WHO (2017) *Global tuberculosis report. Geneva: World Health Organization*.

567 Wirth, T. *et al.* (2008) 'Origin, Spread and Demography of the Mycobacterium tuberculosis
568 Complex', *PLoS Pathog*, 4(9), p. e1000160. doi: 10.1371/journal.ppat.1000160.

569 Yang, C. *et al.* (2012) 'Mycobacterium tuberculosis Beijing Strains Favor Transmission but Not
570 Drug Resistance in China', *Clinical Infectious Diseases*, 55(9), pp. 1179–1187. doi:
571 10.1093/cid/cis670.

572 Yu, Y. *et al.* (2015) 'RASP (Reconstruct Ancestral State in Phylogenies): A tool for historical
573 biogeography'. doi: 10.1016/j.ympev.2015.03.008.

574 Yuan, W. (2006) *La Chine et l'Afrique : [1956-2006]*. [S. l.]: China Intercontinental Press.

575

Data Availability Statement

The additional datasets i.e. additional xml files and scripts for this study can be found here https://github.com/SwissTPH/TBRU_L2Africa.git

Figure Legends

Figure 1. Global phylogeny and geographical distribution of L2 strains. **(A)** Geographical origin (according to United Nations geoscheme) for the 817 L2 strains. The geographical origins with less than 20 strains are colored dark gray and those with missing data light gray. **(B)** Maximum likelihood phylogeny inferred from 33,776 variable single nucleotide positions of the 817 strains. Taxa are colored according to the geographical origin of the strains and the clades are highlighted according to previously defined sublineages. L2 defining markers i.e. deletions (RD) are also mapped onto the phylogeny.

Figure 2. Population structure of L2 strains. Frequency in proportions of L2 sub-lineages across seven geographical regions.

Figure 3. Genetic diversity of L2 strains within geographical regions. **(A)** Nucleotide diversity (π) per site of L2 strains by geography. Error bars are the 95% confidence intervals. **(B)** Pairwise genetic SNP distance of L2 by geography (p values were obtained from Wilcoxon rank-sum tests). Each box represents the 25% and 75% quartiles and the line denotes the median.

Figure 4. Stochastic mapping of the geographic origin on L2 phylogeny. **(A)** Maximum likelihood phylogeny of the 817 MTB Lineage 2 strains. Branches are colored according to their geographical region inferred from stochastic mapping of geographic origin of L2 strains onto the phylogeny. The 13 migration events to Africa (M1–M13) are indicated. (**B**) Proposed scenario for the multiple introduction events of L2–Beijing into Africa. **(C)** Plot summarizing the number introduction events to Eastern and Southern Africa from East- and Southeast Asia.

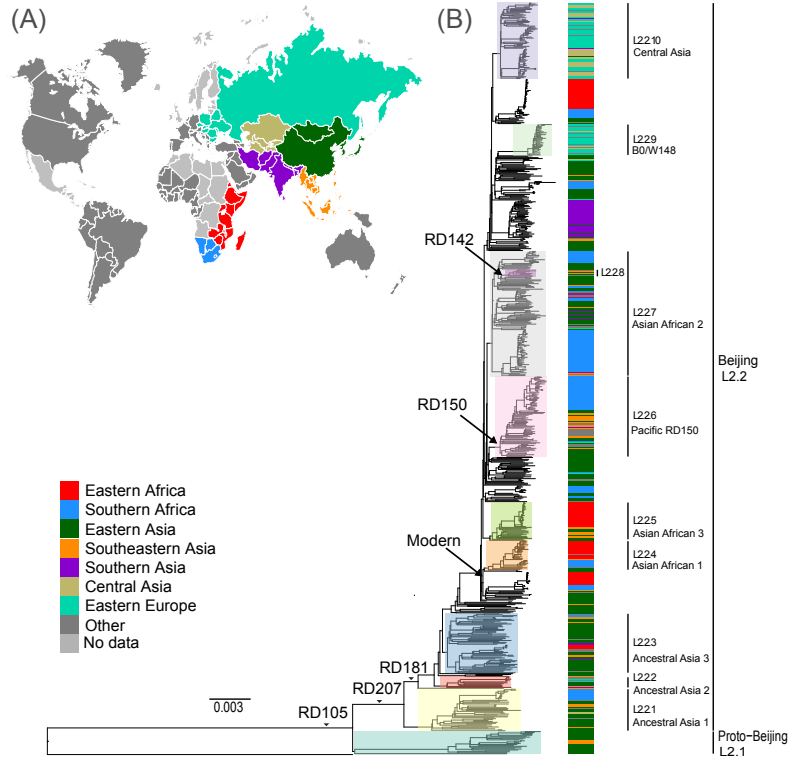
Figure 5. Estimated time in median ages for the introductions of African L2–Beijing (M1-M3 and M6-M13). Introductions to Eastern Africa are colored in red and those to Southern Africa in blue. Migration M10 contained L2–Beijing from both Southern and Eastern Africa. Dotted line marks the year of first anti-TB drug discovery (1943). The error bars correspond to the 95% HPD.

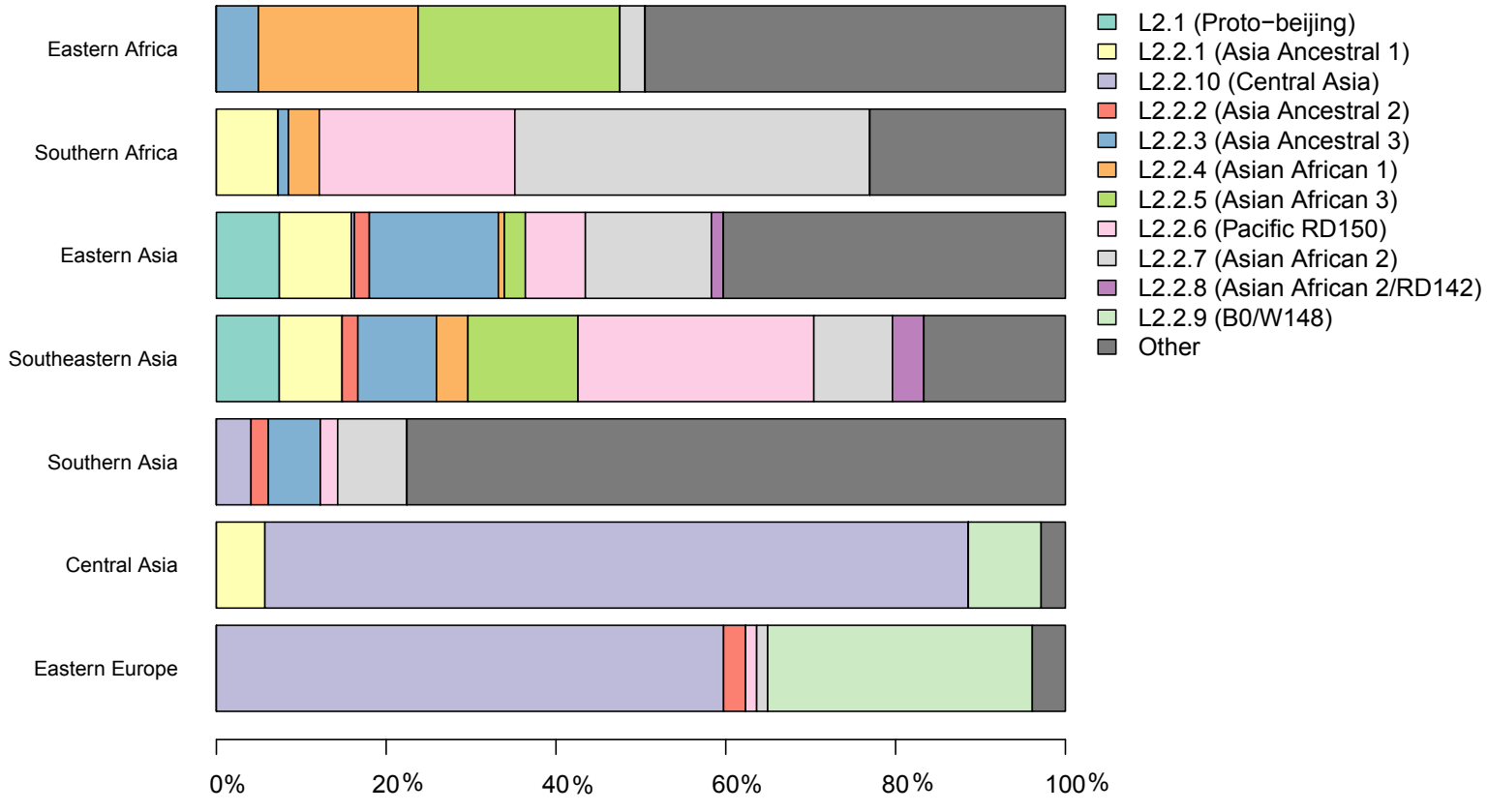
Figure 6. Proportions of drug resistance profiles for L2 strains in seven geographical regions.

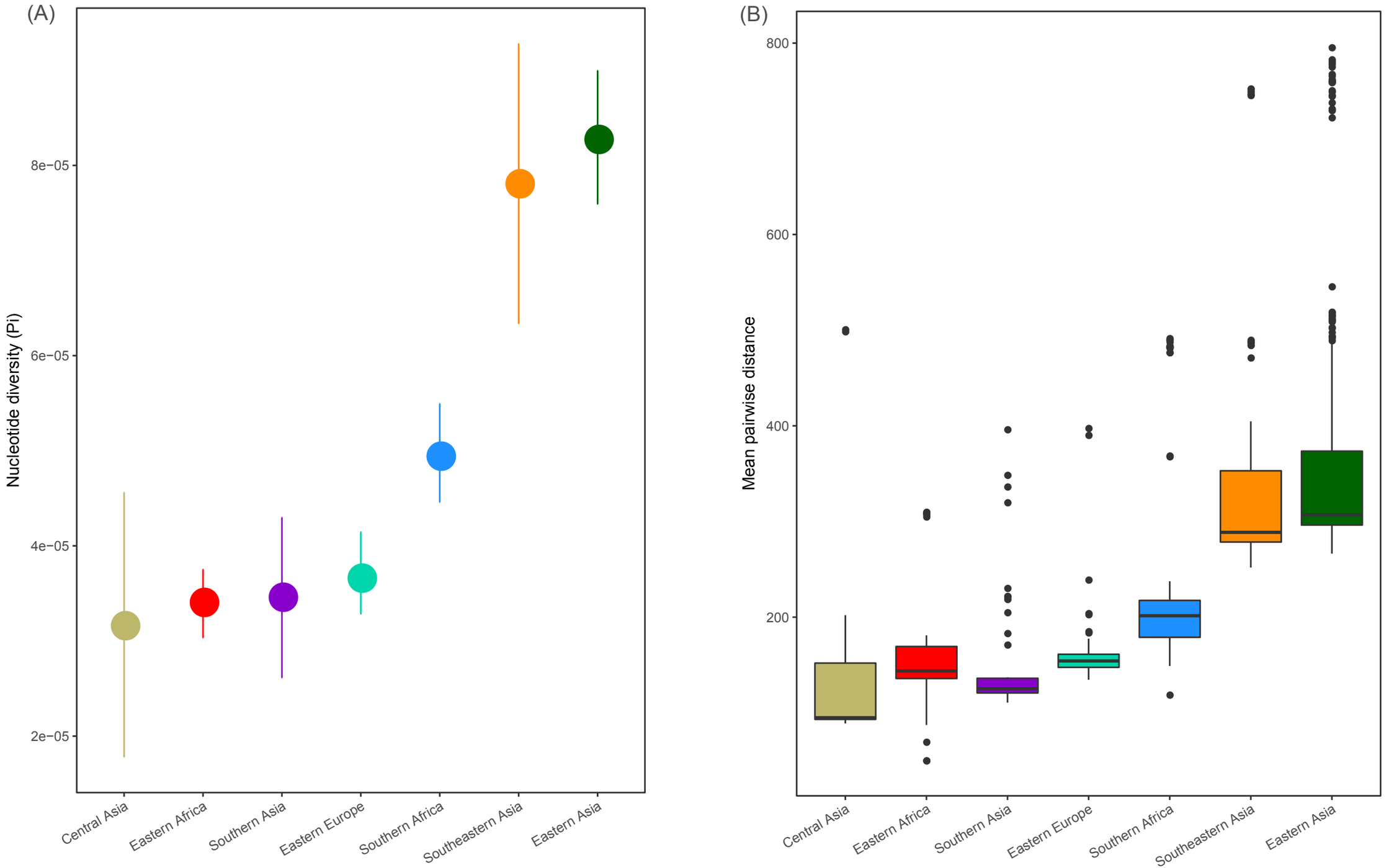
Tables

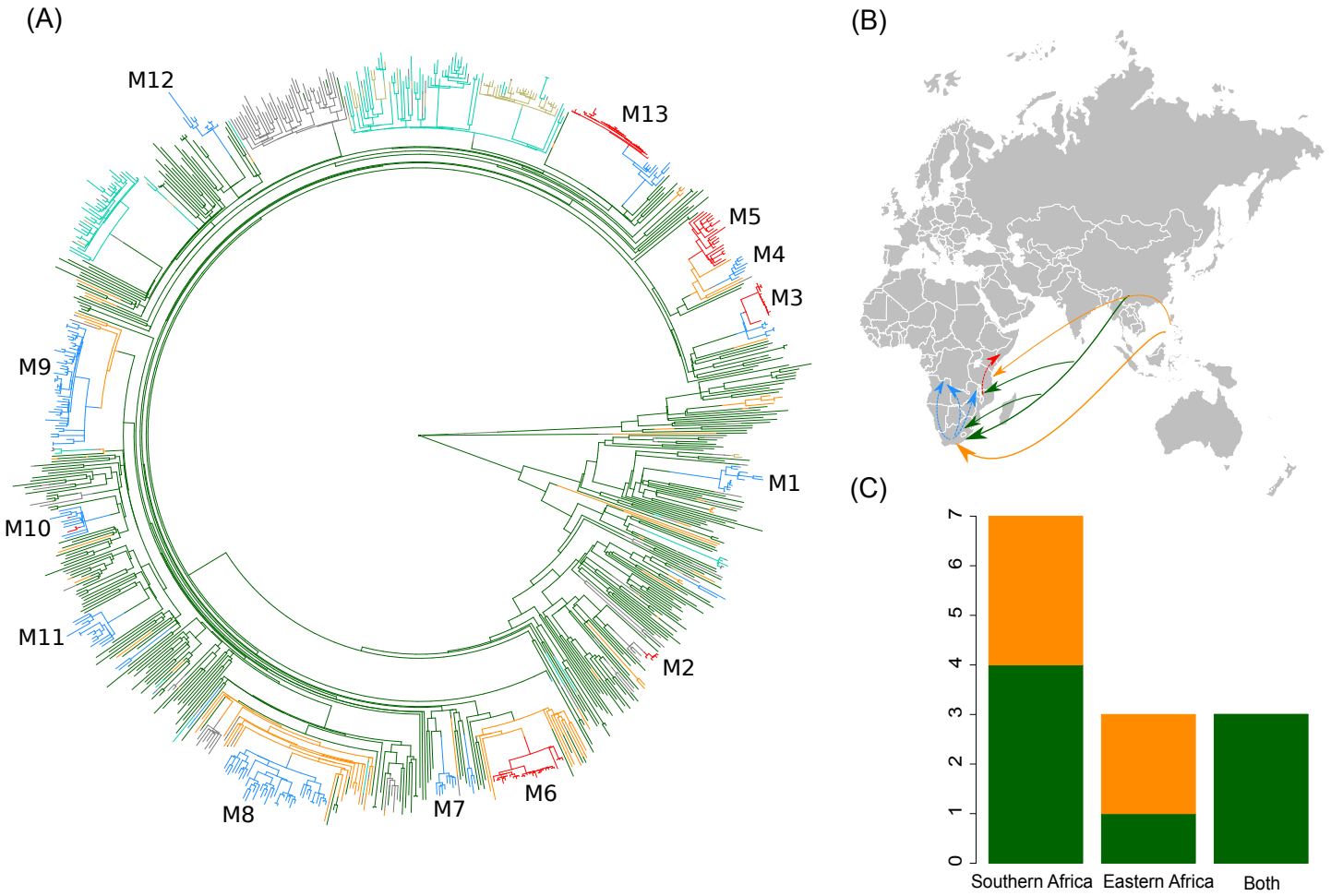
Table 1: Model selection based on path sampling Log-Marginal Likelihood

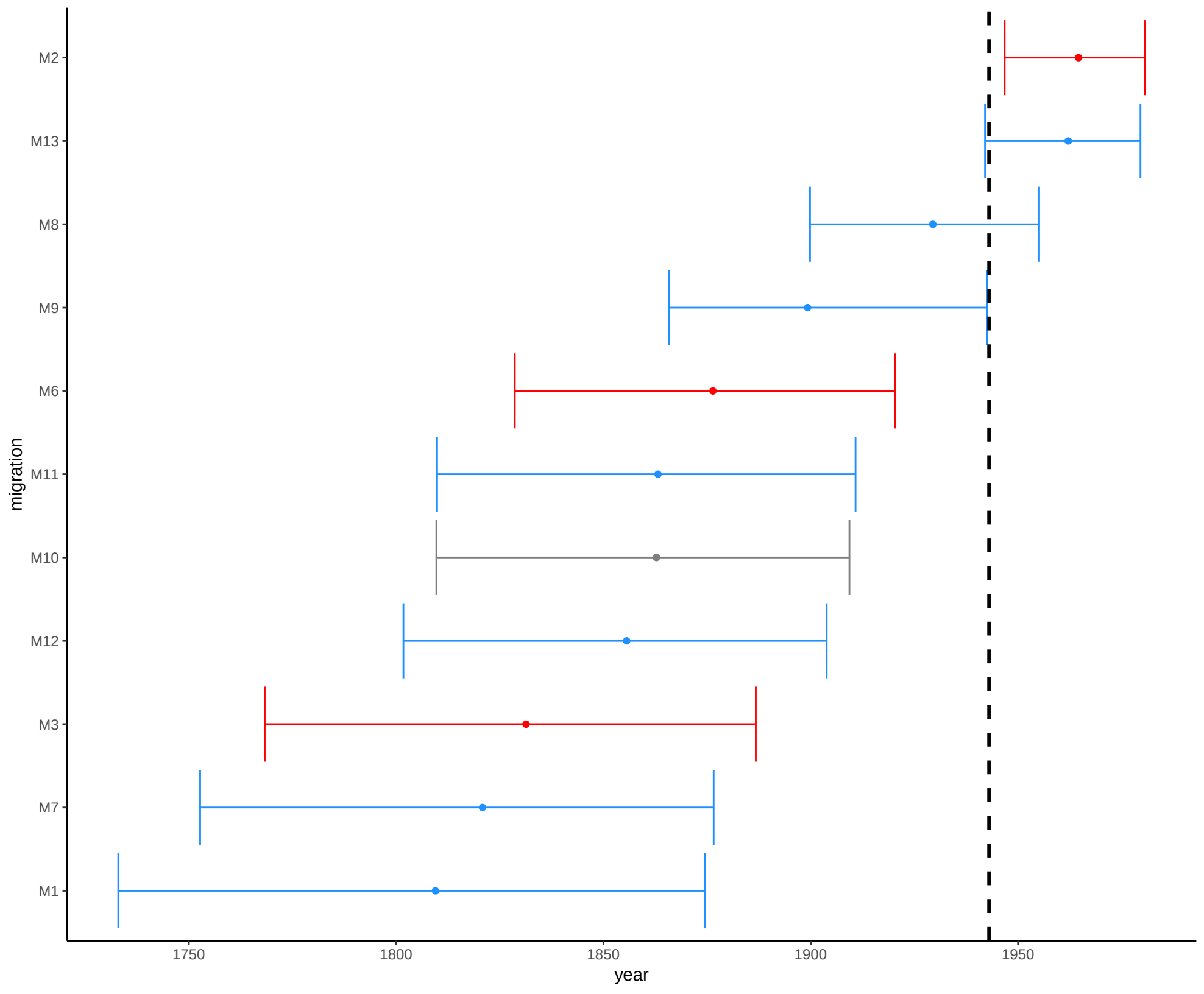
Clock model	Log(e) Marginal Likelihood	Bayes factor (UCLD vs model)
UCLD	-5374827	
Strict	-5374854	27
UCED	-5374897	70

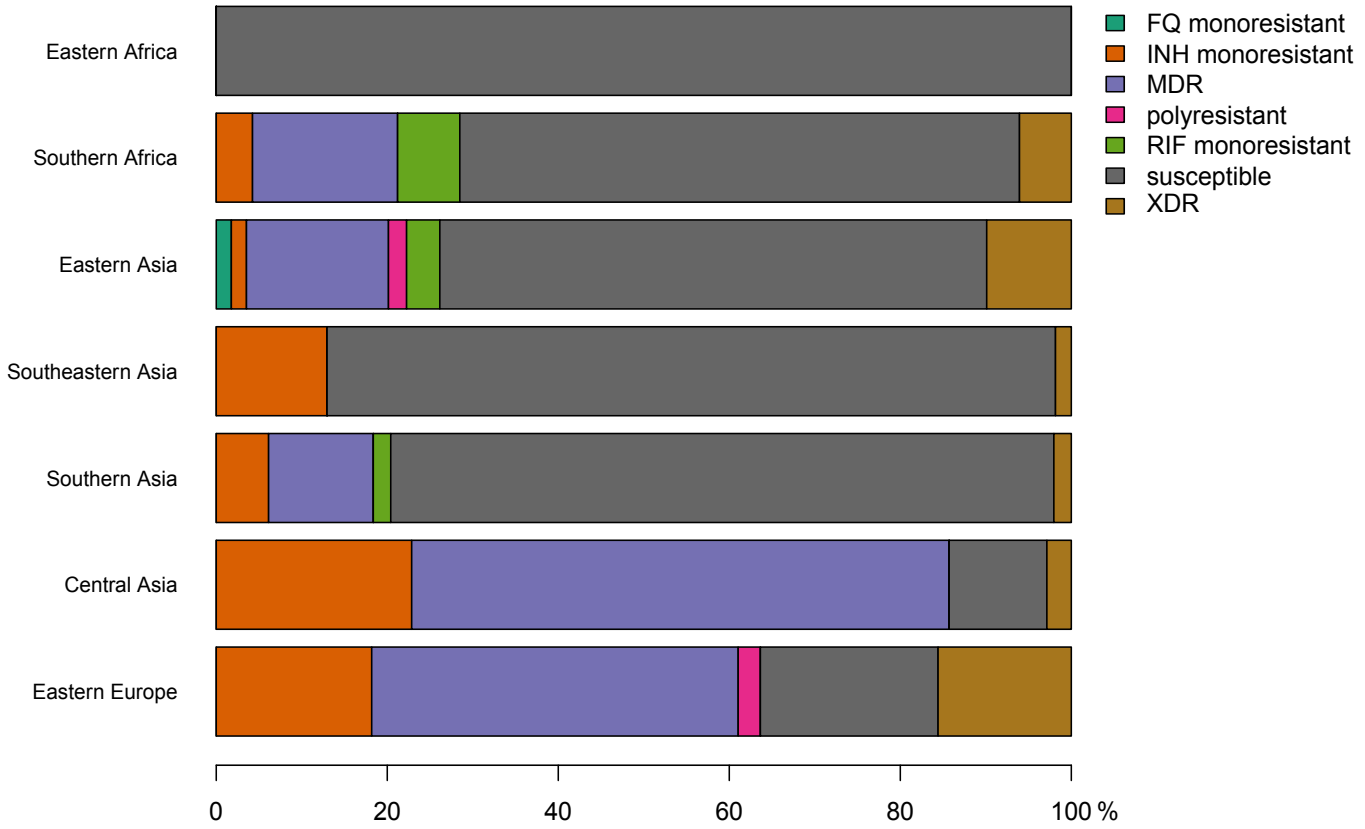

(A)
(B)
Eastern Africa
Southern Africa
Eastern Asia
Southeastern Asia
Southern Asia
Central Asia
Eastern Europe
Other
No data
RD142
RD150
Modern
RD181
RD207
RD105
0.003
L2210
Central Asia
L229
B0/W148
L228
L227
Asian African 2
L226
Pacific RD150
L225
Asian African 3
L224
Asian African 1
L223
Ancestral Asia 3
L222
Ancestral Asia 2
L221
Ancestral Asia 1
Beijing
L2.2
Proto−Beijing
L2.1


Eastern Africa
Southern Africa
Eastern Asia
Southeastern Asia
Southern Asia
Central Asia
Eastern Europe
0%
20%
40%
60%
80%
100%
L2.1 (Proto−beijing)
L2.2.1 (Asia Ancestral 1)
L2.2.10 (Central Asia)
L2.2.2 (Asia Ancestral 2)
L2.2.3 (Asia Ancestral 3)
L2.2.4 (Asian African 1)
L2.2.5 (Asian African 3)
L2.2.6 (Pacific RD150)
L2.2.7 (Asian African 2)
L2.2.8 (Asian African 2/RD142)
L2.2.9 (B0/W148)
Other


(A)
Nucleotide diversity (Pi)
8e−05
6e−05
4e−05
2e−05
Central Asia
Eastern Africa
Southern Asia
Eastern Europe
Southern Africa
Southeastern Asia
Eastern Asia
(B)
Mean pairwise distance
800
600
400
200
Central Asia
Eastern Africa
Southern Asia
Eastern Europe
Southern Africa
Southeastern Asia
Eastern Asia


(A)
M12
M13
M5
M4
M3
M9
M1
M10
M11
M2
M8
M7
M6
(B)
(C)
0
1
2
3
4
5
6
7
Southern Africa
Eastern Africa
Both


M2
M13
M8
M9
M6
M11
M10
M12
M3
M7
M1
migration
1750
1800
1850
1900
1950
year


Eastern Africa
Southern Africa
Eastern Asia
Southeastern Asia
Southern Asia
Central Asia
Eastern Europe
0
20
40
60
80
100 %
FQ monoresistant
INH monoresistant
MDR
polyresistant
RIF monoresistant
susceptible
XDR