Armed conflict and population displacement as drivers of the evolution and dispersal of *Mycobacterium tuberculosis*

Eldholm V*[1], Pettersson J H-O[1], Brynildsrud O[1], Merker M[2], Kitchen A[3], Niemann S[2,4], Rasmussen M[5], Lillebaek T[5], Rønning JO[1], Crudu V[6], Mengshoel AT[1], Debech N[1], Alfsnes K[1], Bohlin J[1], Pepperell CS[7], Balloux F[8]

***Corresponding author: Vegard Eldholm, +4741104428, v.eldholm@gmail.com**

1 Infection Control and Environmental Health, Norwegian Institute of Public Health, Lovisenberggata 8, 0456 Oslo, Norway

2 Molecular and Experimental Mycobacteriology, Forschungzentrum Borstel, Parkallee 1, 23845 Borstel, Germany

3 Department of Anthropology, University of Iowa, Iowa City, USA

4 German Center for Infection Research (DZIF), Partnersite Hamburg-Lübeck-Borstel, Germany

5 Statens Serum Institut, International Reference Laboratory of Mycobacteriology, 5 Artillerivej, DK-2300 Copenhagen, Denmark

6 Microbiology and Morphology Laboratory, Phthisiopneumology Institute, Chisinau, Moldova

7 University of Wisconsin-Madison, School of Medicine and Public Health, Departments of Medicine (Infectious Diseases) and Medical Microbiology and Immunology, Microbial Sciences Building, 1550 Linden Drive, Madison, USA

8 UCL Genetics Institute, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK

Abstract

The 'Beijing' *Mycobacterium tuberculosis* (*Mtb*) Lineage 2 (L2) is spreading globally and has been associated with accelerated disease progression and increased antibiotic resistance. Here we performed a phylodynamic reconstruction of one of the L2 sublineages, the Central Asian Clade (CAC), which recently spread to Western Europe. We find that recent historical events have contributed to the evolution and dispersal of the CAC: our timing estimates indicate the clade was likely introduced to Afghanistan during the 1979 Soviet invasion and spread further following population displacement in the wake of the American invasion in 2001. We also find that drug resistance mutations accumulated on a massive scale in *Mtb* isolates from former Soviet republics following the fall of the Soviet Union, a pattern that was not observed in CAC isolates from Afghanistan. Our results highlight the detrimental effects of political instability and population displacement on tuberculosis (TB) control and demonstrate the power of phylodynamic methods for understanding bacterial evolution in space and time. Although, we did not attempt to reconstruct the age of *Mtb* or L2 as a whole, our dated CAC phylogeny reaches far enough into the past to question the validity of an ancient 'out-of-Africa' origin for *Mtb*.

Keywords:

Mycobacterium tuberculosis, evolution, antibiotic resistance, tip-dating

Significance statement (120 words max)

We employed population genomic analyses to reconstruct the history of dispersal of a major clade of *Mycobacterium tuberculosis* in Central Asia and beyond. Our results indicate that the fall of the Soviet Union and the ensuing collapse of public health systems led to a rise in *M. tuberculosis* drug resistance. We also show that armed conflict and population displacement have aided the dispersal of the clade out of Central Asia via war-torn Afghanistan.

INTRODUCTION

The *Mycobacterium tuberculosis* complex (MTBC) comprises seven main lineages. Of these, lineages 2, 3 and 4 are found across most of the globe but their regional distribution varies and reflects historical and recent human population movements. Lineage 4, the most widely distributed lineage, is spread across Europe, Africa, and the Western Hemisphere, most likely resulting from European colonial history, slave trade and migration. L2 ('L2' and 'Beijing lineage' is used interchangeably throughout the text) has a South East (1) or East Asian (2) origin and has received considerable attention as it is spreading globally (3), might be associated with accelerated progression of disease (4, 5) and is associated with increased antibiotic resistance (5). It has also been suggested that L2 displays an elevated mutation rate relative to other *Mtb* lineages, but studies have yielded differing results in this regard (6, 7).

There is no consensus in the literature on the age of the MTBC and its main lineages and different studies have tried to answer this question using different strategies. One such approach (the 'out of Africa' hypothesis) is based on the assumption of co-divergence of *Mtb* with its human host (1, 8), and suggested that the most recent common ancestor (MRCA) of *Mtb* existed about 40-70 K years ago with the bacillus subsequently spreading globally with human migrations out of Africa (9, 10). By contrast, the two studies that have relied on genomic sequence data using ancient DNA (aDNA) analysis point to a ten times younger origin, around 6,000 years ago (11, 12). Even though calibration with aDNA is becoming the gold standard for dating old evolutionary events, it should be noted that only few non-contemporaneous MTBC genomes are available. One study relied on ~1,000 year-old *M. pinnipedii* isolates, an animal MTBC strain (11). A second study relied on *Mtb sensu stricto* genomes for calibration, but the isolates were only 200-250 years old (12). These two studies yielded similar rate estimates, despite the fact that they included data from very different time periods. The substitution rate estimates of $\sim 5 \times 10^{-8}$ substitutions/site/year (s/s/y) obtained in these aDNA studies are slightly lower than estimates from epidemiological studies and other studies based on contemporaneous sampling, all of which produced rate estimates around 1x10-7 s/s/y corresponding to 0.3-0.5 substitutions/genome/year (6, 13-18).

The origin and spread of the Beijing lineage has also been vigorously debated. According to a recent phylogeographic analysis of L2 genomes, the lineage emerged in South East Asia some 30 K years ago, and subsequently spread to Northern China where it experienced a massive population expansion, purportedly related to the Neolithic expansion of the Han Chinese population (1). The 30 K age was obtained by extrapolating from the aforementioned 70 K age for the MTBC. Another attempt to reconstruct the age and evolutionary history of L2 and its clonal complexes (CCs), based on a massive global collection of Mycobacterial Interspersed Repetitive Unit (MIRU) genotyping data complemented with genome sequencing, resulted in an age of about 6.6 K years for the whole lineage and about 1.5-6 K years for each of the CCs (2). However, this study also relied on strong assumptions in particular concerning the underlying mutation model and mutation rate of the MIRU markers (2, 10).

Until recently, fine-scaled phylodynamic and phylogeograpic methods were mainly applied to rapidly evolving taxa, such as RNA viruses (19). The increased availability of whole-genome sequences has shifted the limits of what can be regarded as measurably evolving pathogens to also include bacteria (20) including *Mtb* (13, 21) despite its relatively slow substitution rate compared to most other bacterial pathogens (22). Here, we apply phylodynamic methods, calibrated with sampling dates (tip-dating), to a collection of *Mtb* isolates from Europe, South and Central Asia. The isolates belong to a L2 clade we term the Central Asian Clade (CAC). The CAC corresponds to the MIRU-defined CC1 (2) and includes the Russian Clade A (23). The isolates included in the study cover a sampling period of 15 years, and even though we did not attempt to reconstruct the age of *Mtb* or L2 as a whole, our dated CAC phylogeny reaches far enough into the past to question the validity of the ancient 'out-of-Africa' scenarios for *Mtb*.

We also show that the evolution and dispersal of the CAC in Eurasia have been shaped by identifiable recent historical events. Specifically, we find that being an ex-Soviet state is a major risk factor for relative multidrug-resistant TB (MDR-TB) prevalence globally and that this pattern holds true within the CAC. We were able to trace the introduction of this clade to Afghanistan around the 1979 Soviet invasion and document its subsequent spread across

Europe following migration events in the wake of recent armed conflict. Our results highlight the detrimental effects of political instability and population displacement for global TB control and demonstrate the power of phylodynamic methods for understanding bacterial evolution in time and space.

RESULTS AND DISCUSSION

Defining the Central Asian Clade

In order to investigate the recent history and spread of an *Mtb* L2 clade associated with Afghan refugees in Norway, *Mtb* genomes from a recent large TB outbreak mainly affecting Norwegian and Afghan nationals in Oslo, Norway (Norheim et al, in review J Clin Microbiol) were included in the study together with related isolates from Norway, Denmark, Germany and Moldova. In addition, we included sequencing data from other relevant studies (see Materials and Methods). A whole-genome SNP phylogeny was constructed as described in the materials and methods section. From this phylogeny it was clear that the Oslo outbreak belongs to a relatively diverse Afghan strain family (Fig. 1A, orange highlighting). This Afghan strain family belongs to a larger clade that includes the previously described Clade A from Russia (23) and Central Asian isolates from a recent global study (2) (Fig. 1, blue highlighting). Interestingly, Casali and colleagues noted that Clade A isolates were consistently found at a higher frequency east of the Volga whereas the other dominant clade in Russia, Clade B was more frequent west of the river (23). We therefore term this clade, encompassing both clade A and Central Asian isolates as defined in earlier studies (2, 23), the Central Asian Clade (CAC) (Figure 1A).

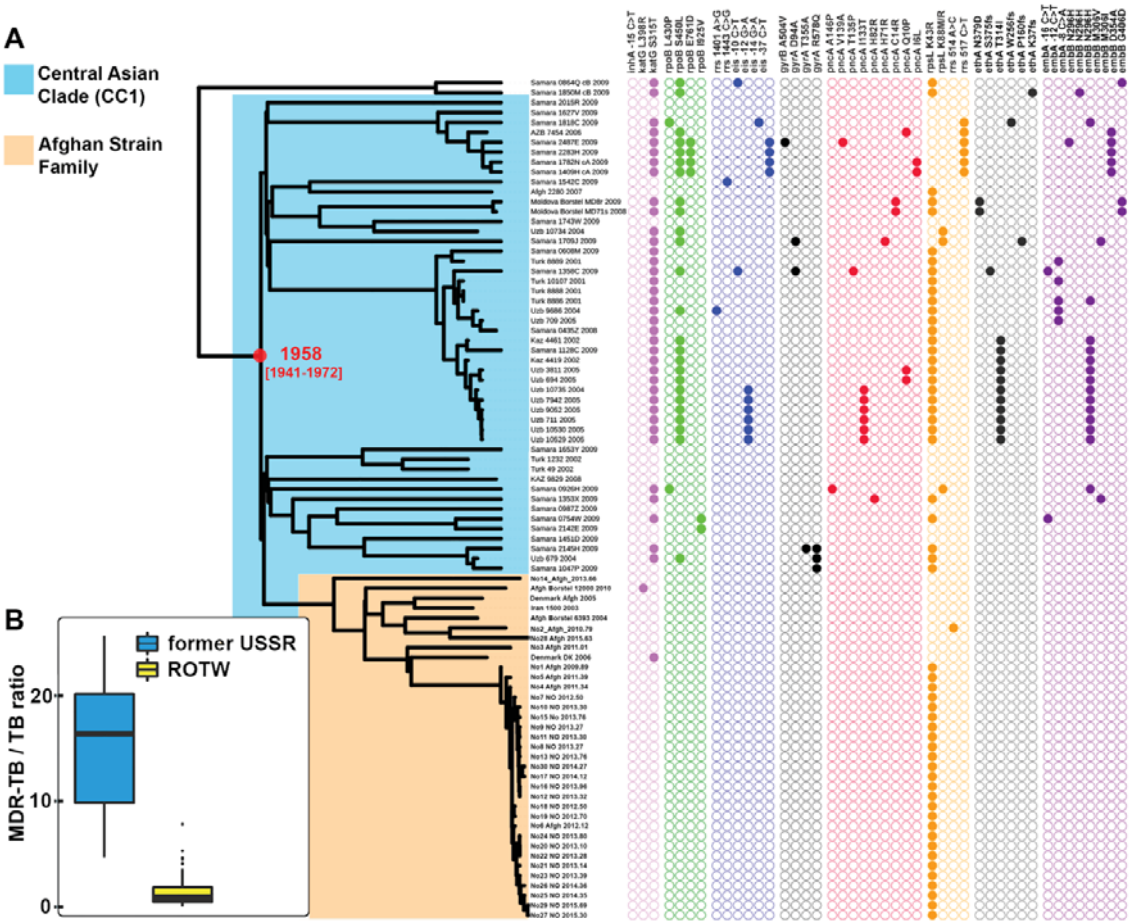


Figure 1. **Phylogenetic placement and antibiotic resistance of *Mtb* isolates in the study.** (A) Bayesian dated phylogeny of the Central Asian Clade (CAC). The Afghan strain family and the Central Asian Clade to which it belongs are highlighted in orange and blue respectively. Filled dots indicate the presence of mutations colored by the compound to which they are known or predicted to confer resistance (magenta: isoniazid, purple: rifampicin, blue: kanamycin, green: fluoroquinolones, yellow: pyrazinamide, orange: streptomycin, red: ethionamide, grey: ethambutol). The age of the CAC most recent common ancestor (MRCA) is indicated in red. Two clade B isolates (23) were used as outgroup. (B) Relative prevalence of multidrug-resistant TB (MDR-TB) stratified by a history of Soviet Union allegiance (blue: ex-Soviet states, yellow: rest of the world).

The fall of the Soviet Union and the rise of MDR-TB

Mapping of known and putative resistance mutations on the phylogeny revealed that isolates originating in Central Asia were strongly enriched in resistance mutations relative to Afghan isolates (Fig. 1A). The countries in Central Asia were all part of the Soviet Union until its fall in 1991. To investigate geographic patterns of drug resistance in more detail, we divided countries into two groups: ex-Soviet states and the rest of the world (ROTW) and analyzed global data on relative prevalence of MDR-TB (*Mtb* resistant to first-line drugs isoniazid and rifampicin). Even though it is widely acknowledged that MDR-TB represents a particularly acute problem in many ex-Soviet countries, the strength of the association we find remains striking (Fig. 1B, Wilcoxon Rank Sum Test: $p<0.001$, $W=2577$). To examine in more detail whether our CAC data supported a role of the fall of the Soviet Union in the rise of resistance within the clade, we mapped individual resistance mutations to nodes in the dated phylogeny. From this phylogeny it is clear that the majority of transmitted resistance mutations evolved in the years following the collapse of the Soviet Union (Fig. S1). Together, these findings support the notion that external factors, namely the fall of the Soviet Union and the ensuing breakdown of public health systems, rather than features specific to the Beijing lineage, are to blame for the extreme rates of drug resistance in parts of the region.

A *recent origin of the Central Asian Clade*

To investigate the temporal evolution and spread of the CAC and the Afghan strain family in detail, we performed Bayesian phylogenetic analyses using BEAST 1.7.4 (24) with tip-dates (sampling dates) for temporal calibration. We investigated root-to-tip distances as a function of sampling time and employed tip-randomization to assess the strength of the temporal signal in the data (see materials and methods). Both tests revealed a strong temporal signal in the data. Bayesian phylogenetic analyses under different clock and demographic models on various sample subsets, resulted in similar ages of the MRCAs of both the CAC and the Afghan strain family, respectively (table 1).

Table 1. Estimated time to most recent common ancestor (TMRCA) for the Central Asian clade (CAC) and the Afghan strain family (ASF)

Sample set	demographic model	TMRCA [95% HPD] Central Asian clade	TMRCA [95% HPD] Afghan strain family
ASF	Skyride	na	1978 [1963–1990]*
ASF	Skyride (RC)	na	1980 [1967–1992]
ASF	Constant	na	1973 [1951–1989]
ASF	Exponential	na	1978 [1965–1989]
ASF	Expansion	na	1979 [1964–1992]
ASF	Logistic	na	1972 [1948–1990]
CAC	Skyride	1958 [1941–1972]*	1972 [1958–1985]
CAC	Skyride (RC)	1959 [1942–1974]	1971 [1957–1984]
CAC representatives[#]	Skyride	1951 [1921–1975]	1968 [1947–1986]
CAC representatives[#]	Skyride (RC)	1955 [1920–1981]	1967 [1940–1989]
CAC (÷ Samara)	Skyride	1941 [1914–1964]	1960 [1940–1979]

Strict clock used unless otherwise specified. RC= relaxed clock, HPD = Highest posterior density

*Reported in text

[#] Maximum one isolate included per year per patient country of origin

We estimate time of the MRCA (TMRCA) of the CAC to be 1958 [95% HPD: 1941–1972], which deviates considerably from a previous study based on MIRU data that estimated the age of the Beijing lineage clonal complex 1 (corresponding to the CAC) to be 4,415 [95% HPD: 2,569–7,509] years old (2). In our phylogenetic reconstruction, the CC1 isolates all fall within the CAC and we thus expect TMRCA of the CC1 to be identical or nearly identical to

197 the TMRCA of the CAC. The TMRCA estimates of CC1 were based on a mean MIRU mutation
198 rate per year of 10^{-4} (2, 10). To investigate the mean MIRU evolutionary rate in our samples,
199 we first constructed a tip-dated genome phylogeny including only isolates with available
200 MIRU data (excluding isolates from Samara, Russia). The total branch length of the
201 phylogeny, corresponding to the total evolutionary time (years) elapsed was found to be
202 848 years (95% HPD: 845–852 years). Subsequently we annotated and counted repeat
203 expansion and contraction events (Fig. 2). Only nine of the 24 MIRU loci had undergone any
204 changes in repeat number among the sampled isolates. This corresponds to a mean per-
205 locus MIRU mutation rate of 1.1×10^{-3} mutations per locus per year (Dataset S3), which is
206 about 10-times higher than the rate used as a prior in the previous study. The estimated
207 rate is, however, well in line with other recent rate estimates based on whole genome
208 sequencing of serial *Mtb* isolates from Macaque monkeys and model-based Bayesian
209 estimates (25, 26). Also of note is the number of homoplasies in the MIRU data: out of a
210 total of 23 repeat gain/loss events, seven occurred twice on independent occasions (i.e. on
211 different branches) and thus correspond to homoplasies. That is, 14 of a total of 23 events
212 represented homoplasic events. Furthermore, we observed five occasions of likely
213 simultaneous loss of two repeats, which are more parsimoniously explained by mutations
214 involving two tandems repeats (although stepwise loss in unsampled strains cannot be ruled
215 out). This suggests that MIRU evolution does not follow a strict stepwise mutation model as
216 assumed previously (2). Together, these observations suggest that MIRU data is not an ideal
217 marker for evolutionary inference over long time-scales.

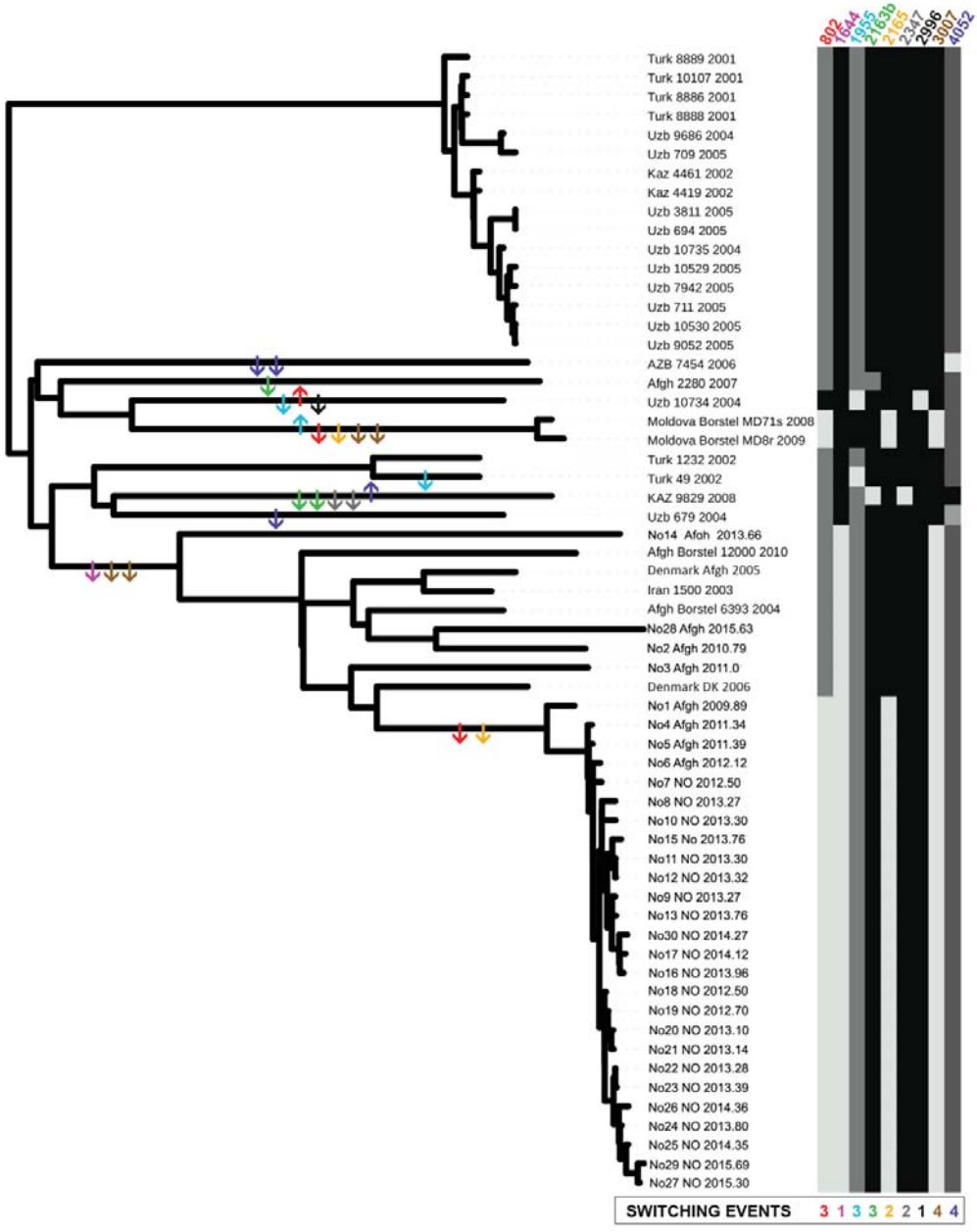
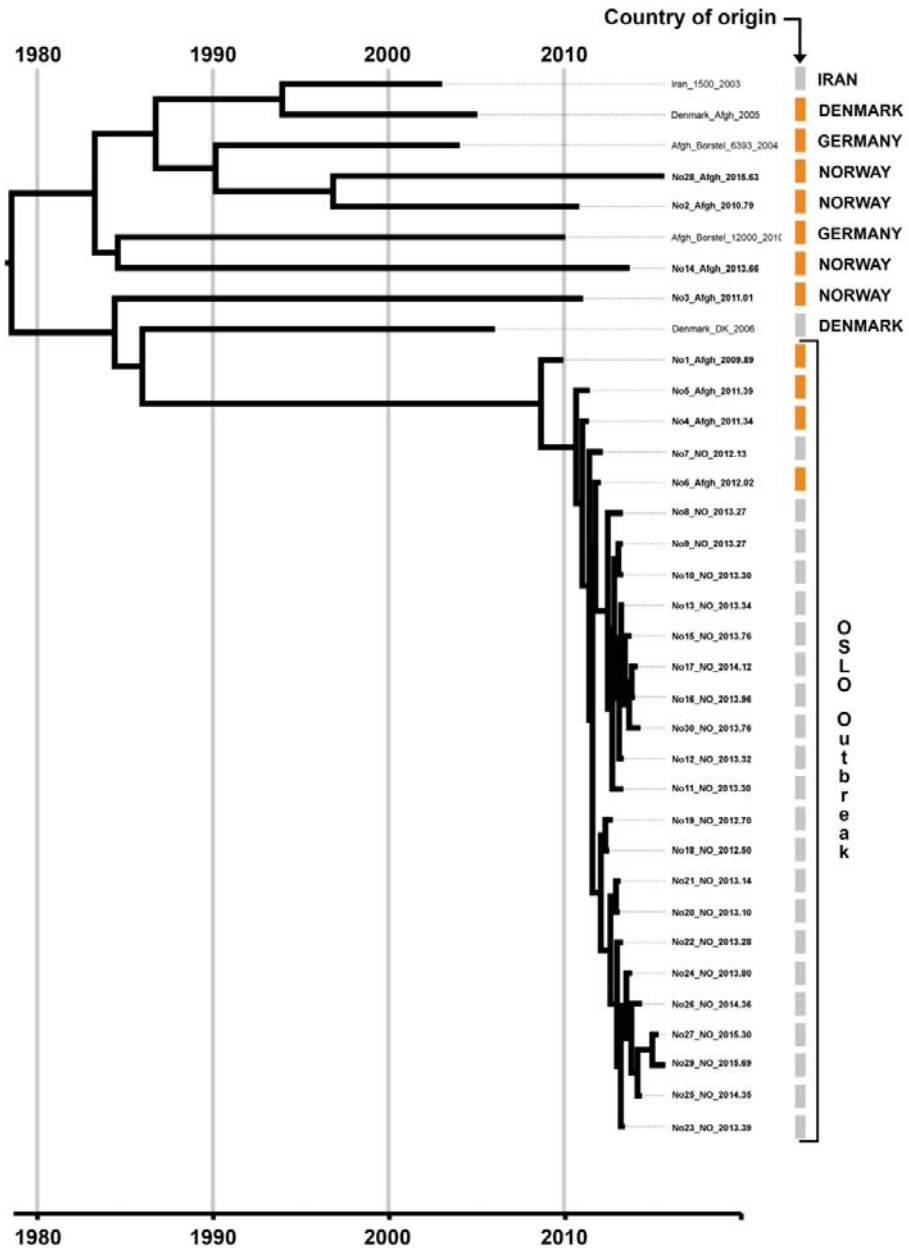


Figure 2. MIRU repeat changes mapped on whole-genome tip-dated phylogeny. Changes in repeat number of nine variable MIRU loci annotated on the right. Individual state change events are indicated by arrows in the phylogeny. The arrows are colored to match the color of individual MIRU loci and the direction of the arrows indicates repeat expansion (up) or contraction (down). The "switching events" box summarizes the number of times individual MIRU loci have added or lost a repeat unit.

The spread of the CAC: the role of armed conflict and population displacement

Our TMRCA estimates suggest that the CAC was introduced to Afghanistan from Soviet Central Asia coincident with the 1979 Soviet invasion of the country (table 1). A dated phylogeny including only isolates belonging to the Afghan strain family revealed that, apart from the Oslo outbreak, individual isolates generally represented isolated TB cases among Afghan refugees in Europe. All cases had been diagnosed between 2003 and 2015 and, again excluding the Oslo outbreak, the isolates were always situated on long terminal branches stretching 10–30 years back in time (Fig. 3). These observations suggest that these TB cases represent multiple individual introductions of the strain to Europe with Afghan refugees in the wake of the continued violent conflicts in the country. The long terminal branches are consistent with reactivation of latent disease in refugees, which in one case was followed by a local outbreak in the receiving country, identifiable by very short terminal branches (Fig. 3).

When interpreting our phylogenetic analyses in the light of historic events in the region, it appears that armed conflict has played a major role both in introducing the CAC to Afghanistan (Soviet invasion) and in the subsequent repeated export of the clade with Afghans fleeing the country in the wake of the American invasion in 2001. A hypothetical scenario for the spread of the CAC and the Afghan strain family in time and space is presented in Fig. 4.



246

247 **Figure 3. Bayesian evolutionary phylogeny of the Afghan strain family**. Colored bars
248 indicate country of origin of the patient: Afghanistan (orange), other countries (grey). The
249 country of isolation is annotated to the right.

250

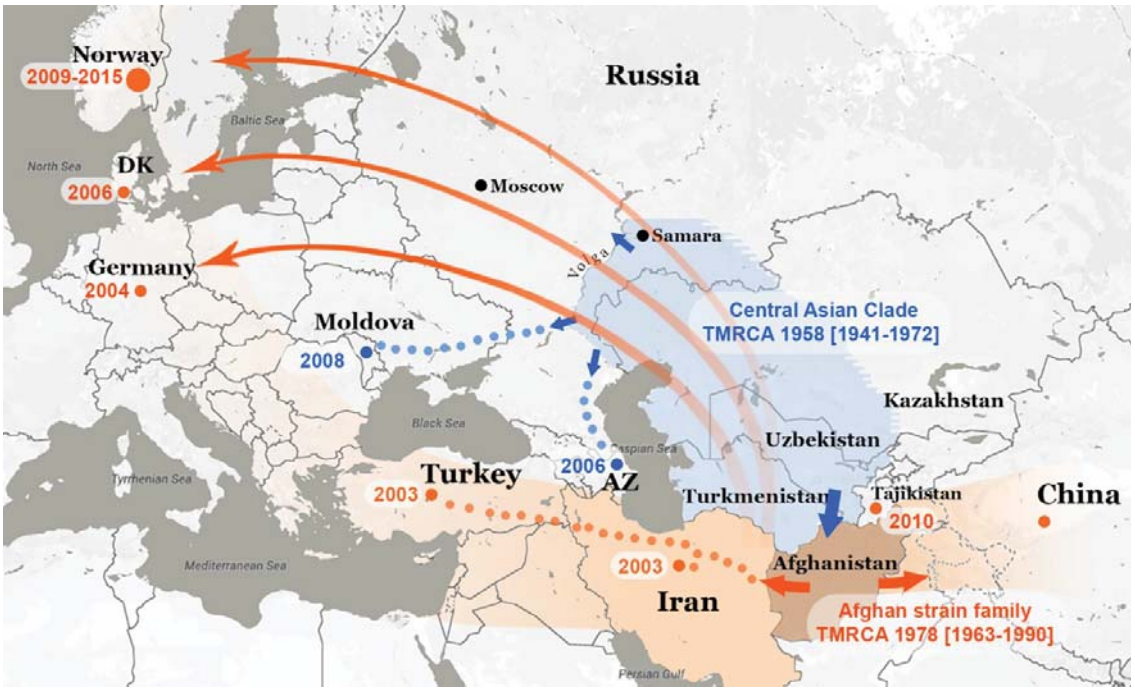


Figure 4. Scenario for the spread of the Central Asian Clade (CAC) and the Afghan Strain family (ASF) in time and space. Based on the origin of sampled patients, the area shaded blue is the heartland of the CAC, whereas shades of orange illustrate the spread of the ASF. Dots represent cases or clusters of cases belonging to either the CAC or the ASF based on genome sequences, except the cases in Turkey, China and Tajikistan for which only MIRU data were available. The sampling year of clinical isolates is provided for each case or cluster of cases.

Substitution rates through time

The origin and subsequent evolutionary history of *Mtb* have been the object of debate (1, 9, 11, 12). It has been suggested that a high degree of congruence between human and *Mtb* phylogenies supports a scenario of co-divergence for the two organisms and that the age of the MRCA of *Mtb* thus mirrors the timing of the migrations of anatomically modern humans out of Africa about 40 K – 70 K years ago (9). However, another study failed to identify such a congruence in phylogenies and did not find support for a co-divergence scenario when employing a host of formal tests (16). Furthermore, the two studies employing aDNA to calibrate MTBC phylogenies both estimate an age of about 6 K years for the TMRCA of extant *Mtb* (11, 12).

We estimated a substitution rate for the CAC of $2.7x10^{-7}$ [95% HPD: $1.3x10^{-7}$ – $3.4x10^{-7}$) s/s/y resulting in a TMRCA estimate of 1958 (95% HPD: 1941–1972]. The age of the Beijing lineage has previously been estimated to about 6 K years (2, 9) or 30 K years (1). Furthermore, the age of a clonal complex corresponding to the CAC (CC1) has been estimated to be about 4.4 K years old (2). The discrepancy between this estimate and the age of about 58 years obtained here by tip-date calibration is striking. However, both root-to-tip analyses and tip date randomization (see materials and methods) suggest that our dating analyses are robust.

The substitution rate estimated for the CAC is slightly higher than previous rate estimates from studies of modern, heterochronous samples, but well within the margin of error for estimates obtained in similar studies (Fig. 5). Interestingly, the other lineage-specific tip-dated rate estimates were all obtained for Lineage 4 isolates, and it is thus possible that the higher rate obtained for the CAC (L2) in the present study, although not significant, might reflect an intrinsically higher mutation rate for L2 lineages (6). The similarity between rates from contemporaneous studies and the two employing aDNA for temporal calibration is also striking even if both *Mtb* aDNA studies point to slightly lower mutation rates. This difference might partly represent time dependency in mutation rate estimates, due to the fraction of slightly deleterious mutations being eliminated over longer periods of time (27). A parallel observation of mutation rate estimates decreasing moderately when older samples are included in the analysis has also been observed in mitochondrial genomes (28) and the agent of the plague, *Yersinia pestis* (29).

This being said, while time-dependency is statistically detectable and likely to be a genuine and general phenomenon, the effect is quantitatively subtle and not compatible with the extreme deceleration in substitution rates over time that would have to be invoked to reconcile these studies with 40-70 K ages for *Mtb* generated under the ancient 'out of Africa' scenarios (9). All current studies based both on ancient and modern samples where mutation rates were directly inferred form the data support the notion that the MRCA of *Mtb* circulating today existed approximately 6 K years ago. This does not rule out that TB is a more ancient disease, as suggested by archeological studies (30, 31). Indeed, the MRCA of

currently extant *Mtb* strains could be younger than TB as a result of a clonal replacement in the global *Mtb* population. It is also possible that the disease resembling TB in the archeological record was caused by an organism other than what is currently identified as *Mtb*.

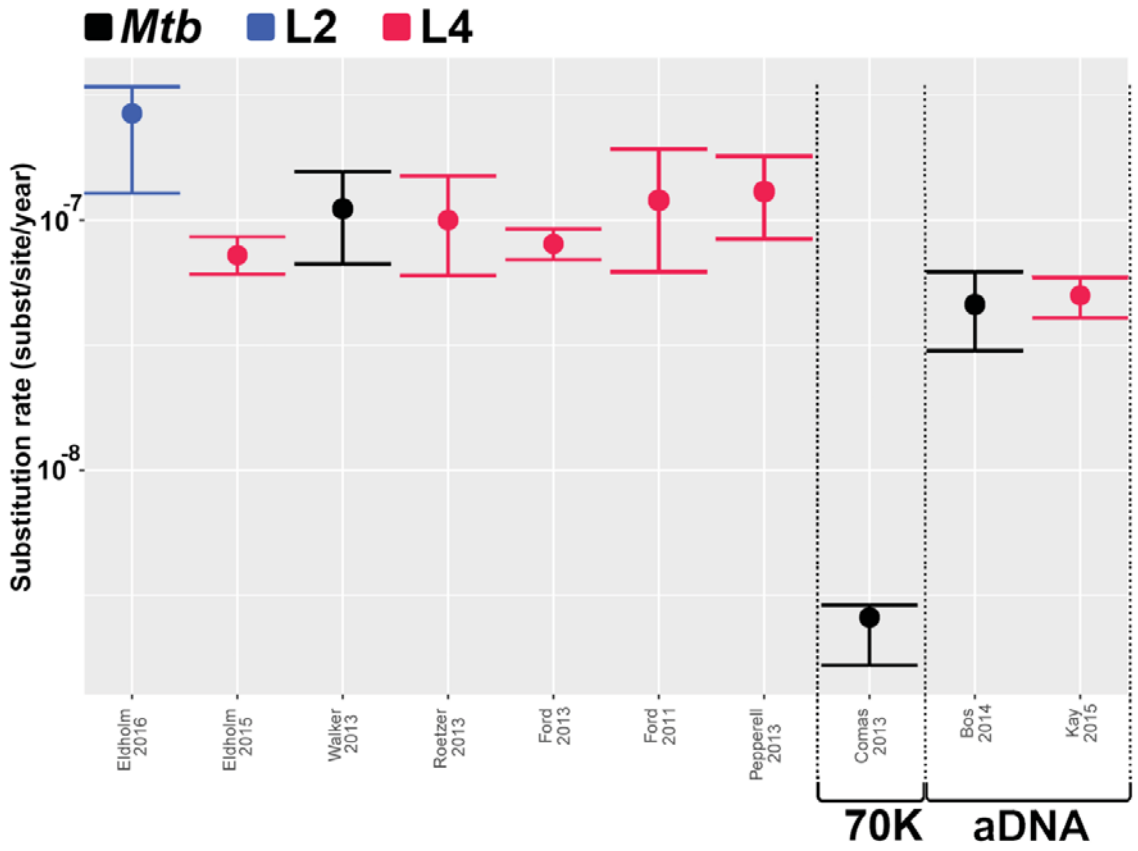


Figure 5. Estimated *Mtb* substitution rates in published datasets. Colors indicates the lineage to which the samples under study belong (Blue: Lineage 2; Red: Lineage 4; Black: all). Studies employing aDNA (Kay 2015 and Bos 2014) and human-*Mtb* co-divergence (Comas 2013) for calibration are annotated separately. The other studies used tip dating (Eldholm 2016, Eldholm 2015, Ford 2013 and Roetzer 2013), historical information (Pepperell 2013) or counted mutations in paired (Walker 2013) or serial isolates (Ford 2011).

MATERIALS AND METHODS

Samples

We included samples from a TB outbreak detected at an Oslo educational institution for young adults in 2013 (Norheim et al, in review J Clin Microbiol) with the last cases belonging to the outbreak diagnosed in 2015. In addition, a search through an in-house database revealed the presence of four *Mtb* isolates from Norway with a MIRU profile (Mtbc15-9 code: 1047-189) that had only two repeat differences from the larger outbreak (Mtbc15-9 code: 10287-189). In total, 26 samples from 24 patients were available from the outbreak (all samples from culture positive patients) and four isolates from the smaller cluster. The earliest cases in the outbreak as well as the four cases in the smaller cluster were all Afghan immigrants to Norway, indicating that these related MIRU types were representatives of a larger reservoir of strains circulating in Afghanistan. To assess whether these two MIRU types were part of one or more larger groups of strains globally, we searched through the MIRU patterns published in a recent extensive global study of L2 isolates [4987 isolates from 99 countries (2)]. We included all sequenced isolates that differed at no more than two MIRU loci from either of the two types described above. As this also included the MIRU type 94-32, making up the majority of CC1, we included all sequenced CC1 isolates from the Merker study (2). An additional four isolates harboring the 1047-189 MIRU pattern and two isolates differing from the 10287-189 pattern at two loci were sequenced for the current study, including five from the global study (2), and one identified in an in-house database at Research Center Borstel, Germany. Finally, a numerically matching sample of genomes from a large genome study centered in Samara Oblast, Russia was included. Included samples can be found under study accessions PRJEB12184, PRJEB9680, ERP006989 and ERP000192. Detailed information on samples included in the study is provided as supplementary datasets S1 and S2.

Calling single nucleotide polymorphisms

Genomic DNA isolation and preparation of sequencing libraries was performed following a published protocol (32) except that we used the Kapa HyperPlus library preparation kit

(KAPA Biosystems, Wilmington, Massachusetts, USA) and its enzymes for DNA fragmentation rather than the Kapa High Throughput Library Preparation Kit. Six-nucleotide barcodes from Bioo Scientific (Bioo Scientific, Austin, Texas, USA) were used for indexing. Illumina raw sequencing reads were mapped against the *M. tuberculosis* H37rv genome (NC_000962.3) using SeqMan NGen (DNASTAR). SNPs in or within 50 bp distance of regions annotated as PE/PPE genes, mobile elements or repeat regions were excluded from all analyses. Heterozygous SNPs that were found at a frequency of 20-80% of reads in at least one isolate were excluded. Finally, for inclusion of SNPs in our downstream analyses, a minimum depth of eight reads in one strain and at least four reads in all strains was required.

Phylogenetic evolutionary inferences

Maximum likelihood phylogenies were constructed from 1,293 concatenated genome-wide SNPs in Seaview (33). The HKY substitution model was chosen based on model testing as implemented in MEGA v5 (34). Divergence times and evolutionary rates were computed from the same alignments using BEAST 1.7.4 (35). The XML-input file was manually modified to specify the number of invariant sites. The SNPs were partitioned into three classes based on functional annotation: intergenic SNPs (class 1), synonymous SNPs (class 2) and non-synonymous + non-coding RNA SNPs (class 3). Phylogenetic trees were visualized using Figtree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree) and ITOL v2 (36).

Assessment of temporal signal and testing of tip-based calibration

To assess the strength of the temporal signal contained in the serial sampling and evaluate if calibrating the Bayesian phylogeny in BEAST using only tip-dates was adequate, we analyzed the root-to-tip distance of our samples as well as various sub-sampling regimes using Path-O-Gen (http://tree.bio.ed.ac.uk/software/ pathogen/). Maximum likelihood trees were computed in SeaView (33) for a number of different sample subsets (described below), all employing a HKY substitution model as described above. As a complementary assessment of

the temporal signal in the data, we performed date randomization on our datasets using a recently developed R package (37). Sampling dates of the genomes were randomly shuffled 20 times and date-randomized data sets were analyzed with BEAST using the same parameters as described below. If the mean estimate of the TMRCA of the isolates obtained from the real data set does not overlap with the 95% highest posterior density intervals of estimates from the date-randomized replicates, the data set can be considered to have sufficient temporal structure and spread (38).

Root-to-tip regression analyses were performed employing both standard least squares regression and MM-type robust regression (39) and revealed a clear temporal signal both within the ASF and the CAC as a whole. To make sure the estimates were not driven by any particular sample subset, we also ran a root-to-tip regression on a subset of samples including a maximum of one sample per year per country of patient origin. The results from all the regression analyses are available as supplementary material (Fig. S2). Date randomization analyses confirmed that there was a strong temporal signal both when including all isolates and when restricting the analyses to the Afghan strain family (Figs S3 and S4).

Molecular dating

Based on model testing of each partition in MEGA v5 (34), a HKY substitution model was chosen for all three partitions in BEAST. The tree was calibrated using tip dates with sampling dates ranging from 2002 to 2015. Tip dates for each *Mtb* genome were specified in years before the present, with 0 being the most recent sampled isolate. We defined uniform prior distributions for the substitution rates (1×10^{-9} – 1×10^{-6} substitutions per site per year). Initial analyses were performed with a Skyride demographic model (40) but we also performed analyses using constant size, logistic growth, expansion growth and exponential growth demographic models.

Posterior distributions of parameters, including divergence times and substitution rates, were estimated using Markov chain Monte Carlo (MCMC) sampling. For each analysis we ran three independent chains consisting of 30–300 million steps, depending on time to convergence, of which the first 10% were discarded as a burn-in. Convergence to the

stationary distribution and sufficient sampling and mixing were checked by inspection of posterior samples (effective sample size >200). Parameter estimation was based on the samples combined from three different chains. The best supported tree was estimated from the combined samples using the maximum clade credibility method implemented in TreeAnnotator (http://beast.bio.ed.ac.uk/treeannotator). BEAST runs were performed with either a strict or a lognormal relaxed clock. Models for clock rate and demographic scenarios were compared in Tracer (http://beast.bio.ed.ac.uk/tracer) using posterior simulation-based analog of Akaike's information criterion (AICM). The Skyride model (40) was found to outperform the other models tested, albeit only marginally in some cases. A relaxed clock model performed slightly better for the CAC as a whole, whereas a strict clock performed marginally better on the ASF isolates alone. As the estimated TMRCAs for both the CAC and ASF differed by no more than two years between the strict and relaxed clock models (table 1), we report the strict clock estimates in the text for simplicity. The Bayesian phylogenetic tree used to date the TMRCA of the CAC is included as supplementary figures annotated with posterior node probabilities (Fig. S5) and individual node ages (Fig. S6). The results from the model testing are summarized in table S1.

Calculating MIRU evolutionary rates

To calculate the yearly rate of MIRU evolution (contractions and expansions), we first constructed a BEAST phylogeny employing a Skyride model and parameters as described above, but excluding all isolates from Samara, as MIRU typing results were not available for these isolates. Note that the exclusion of the Samara isolates resulted in a slightly older TMRCA than that obtained using other sample subsets (table 1). We then extracted the total branch length of the phylogenetic tree using TreeStat (http://tree.bio.ed.ac.uk/software/treestat/). The sum of branch lengths corresponds to the evolutionary time (in years) of every branch from the sampled tips to the MRCA of all the isolates. The number of repeats of each MIRU locus was then manually annotated on the tree (Fig. 3). The total number of state changes over all 24 MIRU loci over the sum of years covered by the tree was then summed assuming a step-wise mode of MIRU evolution (supplementary dataset S3).

Calculating relative MDR-TB prevalence

TB and MDR-TB prevalence data was obtained from the World Health Organization (http://www.who.int/tb/country/data/download/en/). For TB prevalence, data was available for all countries for the year 2013 and point estimates of prevalence by 100 K individuals were retrieved (e_prev_100k).

For MDR-TB prevalence, the data was collected less systematically, and relies on a mix of surveillance, surveys and models. We used the estimated number of MDR-TB cases among all notified pulmonary TB cases (e_mdr_num), expressed as prevalence per 100 K individuals by dividing by country population size estimates from the same source. We calculated the relative proportion of MDR-TB cases by dividing the prevalence of MDR-TB by the prevalence of TB and multiplying this number by 1000.

Acknowledgments

We would like to acknowledge the technical staff at the National Reference Laboratory for Mycobacteria at the Norwegian Institute of Public Health. VE was funded by a postdoctoral fellowship from the Norwegian Research Council (Grant 221562). FB acknowledges support from the ERC (grant ERC260801 – BIG_IDEA), and the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

References

1. Luo T, *et al.* (2015) Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese. *Proceedings of the National Academy of Sciences* 112(26):8136-8141.
2. Merker M, *et al.* (2015) Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nat Genet* 47(3):242-249.
3. European Concerted Action on New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis (2006) Beijing/W genotype Mycobacterium tuberculosis and drug resistance. *Emerg Infect Dis* 12(5):736-743.
4. de Jong BC, *et al.* (2008) Progression to Active Tuberculosis, but Not Transmission, Varies by Mycobacterium tuberculosis Lineage in The Gambia. *Journal of Infectious Diseases* 198(7):1037-1043.
5. Drobniewski F, Balabanova Y, Nikolayevsky V, & et al. (2005) Drug-resistant tuberculosis, clinical virulence, and the dominance of the beijing strain family in russia. *JAMA* 293(22):2726-2731.
6. Ford CB, *et al.* (2013) Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet* 45(7):784-790.
7. Werngren J & Hoffner SE (2003) Drug-susceptible Mycobacterium tuberculosis Beijing genotype does not develop mutation-conferred resistance to rifampin at an elevated rate. *J Clin Microbiol* 41(4):1520-1524.
8. Comas I, *et al.* (2012) Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet* 44(1):106-110.
9. Comas I, *et al.* (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* 45(10):1176-1182.
10. Wirth T, *et al.* (2008) Origin, Spread and Demography of the Mycobacterium tuberculosis Complex. *PLoS Pathog* 4(9):e1000160.
11. Bos KI, *et al.* (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* 514(7523):494-497.
12. Kay GL, *et al.* (2015) Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun* 6:6717.
13. Eldholm V, *et al.* (2015) Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nat Commun* 6:7119.
14. Roetzer A, *et al.* (2013) Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS Med* 10(2):e1001387.
15. Walker TM, *et al.* (2013) Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis* 13(2):137-146.
16. Pepperell CS, *et al.* (2013) The role of selection in shaping diversity of natural M. tuberculosis populations. *PLoS Pathog* 9(8):e1003543.
17. Ford CB, *et al.* (2011) Use of whole genome sequencing to estimate the mutation rate of mycobacterium tuberculosis during latent infection. *Nat Genet* 43(5):482-486.
18. Lillebaek T, *et al.* (2016) Substantial molecular evolution and mutation rates in prolonged latent Mycobacterium tuberculosis infection in humans. *Int J Med Microbiol* *DOI:10.1016/j.ijmm.2016.05.017*.
19. Drummond A, Pybus O, Rambaut A, Forsberg R, & Rodrigo A (2003) Measurably evolving populations. *Trends in Ecology & Evolution* 18:481 - 488.
20. Biek R, Pybus OG, Lloyd-Smith JO, & Didelot X (2015) Measurably evolving pathogens in the genomic era. *Trends Ecol Evol* 30(6):306-313.

21. Didelot X, Gardy J, & Colijn C (2014) Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol* 31(7):1869-1879.
22. Eldholm V & Balloux F (, in press) Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out. *Trends in Microbiology*.
23. Casali N*, et al.* (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 46(3):279-286.
24. Drummond AJ, Suchard MA, Xie D, & Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29(8):1969-1973.
25. Aandahl RZ, Reyes JF, Sisson SA, & Tanaka MM (2012) A Model-Based Bayesian Estimation of the Rate of Evolution of VNTR Loci in <italic>Mycobacterium tuberculosis</italic>. *PLoS Comput Biol* 8(6):e1002573.
26. Ragheb MN*, et al.* (2013) The mutation rate of mycobacterial repetitive unit loci in strains of M. tuberculosisfrom cynomolgus macaque infection. *BMC genomics* 14(1):1-8.
27. Ho SY, Phillips MJ, Cooper A, & Drummond AJ (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol* 22(7):1561-1568.
28. Rieux A*, et al.* (2014) Improved calibration of the human mitochondrial clock using ancient genomes. *Mol Biol Evol* 31(10):2780-2792.
29. Rasmussen S*, et al.* (Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. *Cell* 163(3):571-582.
30. Baker O*, et al.* (2015) Human tuberculosis predates domestication in ancient Syria. *Tuberculosis (Edinburgh, Scotland)* 95 Suppl 1:S4-s12.
31. Lee OY*, et al.* (2012) Mycobacterium tuberculosis complex lipid virulence factors preserved in the 17,000-year-old skeleton of an extinct bison, Bison antiquus. *PLoS One* 7(7):e41923.
32. Eldholm V*, et al.* (2014) Evolution of extensively drug-resistant Mycobacterium tuberculosis from a susceptible ancestor in a single patient. *Genome Biol* 15(11):490.
33. Gouy M, Guindon S, & Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27(2):221-224.
34. Tamura K*, et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731-2739.
35. Drummond A & Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7(1):214.
36. Letunic I & Bork P (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research* 39(suppl 2):W475-W478.
37. Rieux A & Khatchikian C (2015) TipDatingBeast: Using Tip Dates with Phylogenetic Trees in BEAST (R package).
38. Firth C*, et al.* (2010) Using Time-Structured Data to Estimate Evolutionary Rates of Double-Stranded DNA Viruses. *Mol Biol Evol* 27(9):2038-2051.
39. Yohai VJ, Stahel WA, & Zamar RH (1991) A Procedure for Robust Estimation and Inference in Linear Regression. *Directions in Robust Statistics and Diagnostics: Part II*, (Springer New York, New York, NY), pp 365-374.
40. Minin VN, Bloomquist EW, & Suchard MA (2008) Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Molecular Biology and Evolution* 25(7):1459-1471.

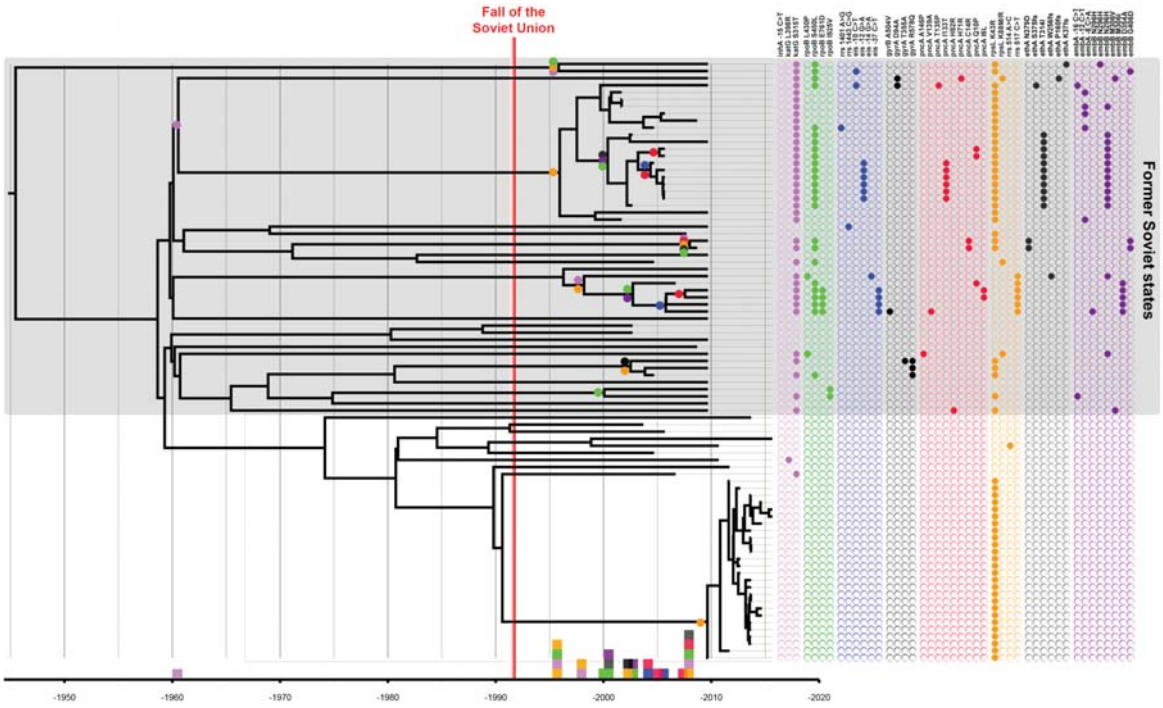


Figure S1. Timed phylogeny with resistance mutations mapped to nodes. Only mutations present in at least two isolates were mapped. The colored boxes at the bottom time-bar indicate the timing of individual mutation events.

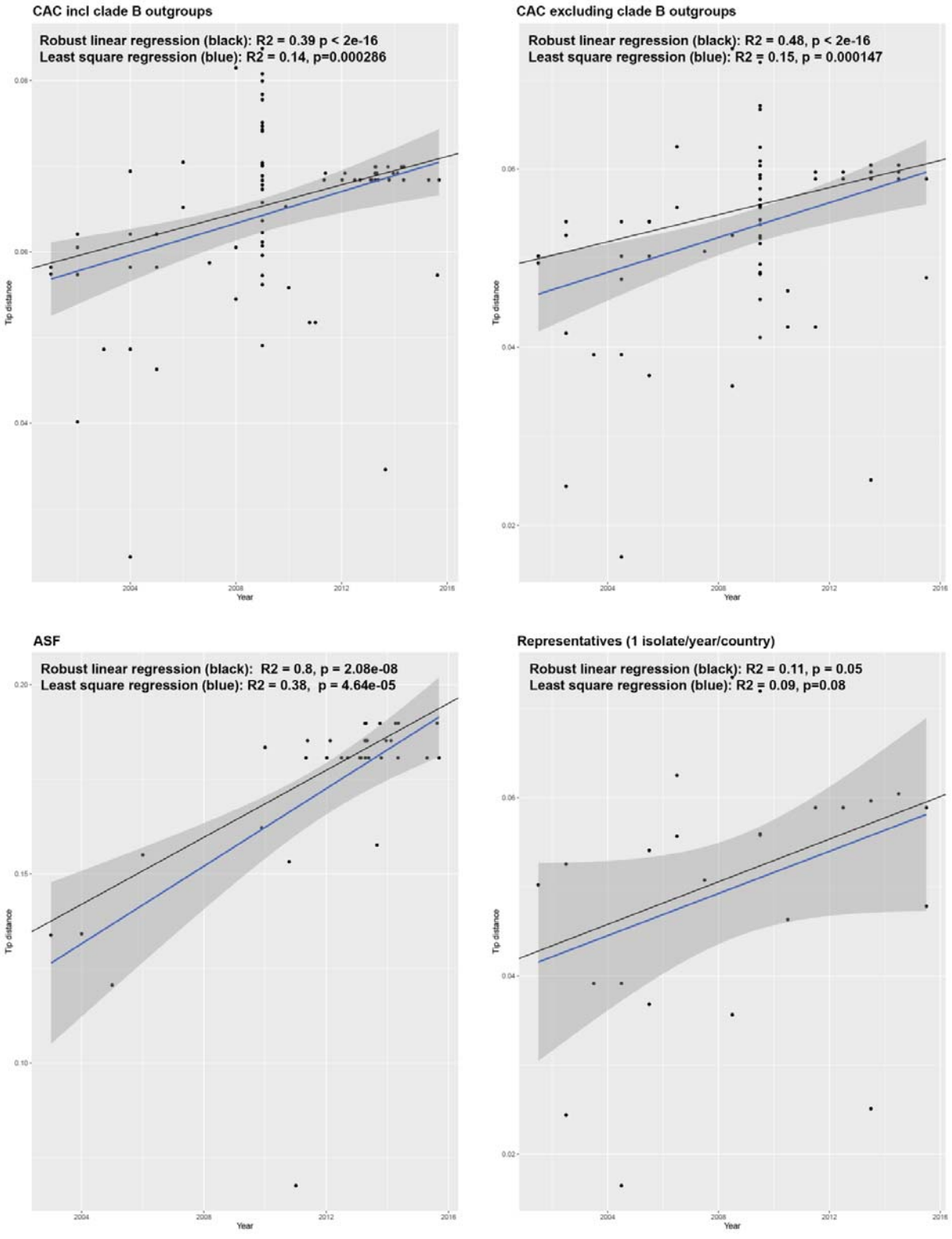


Figure S2. Root-to-tip regression including various sample sets.

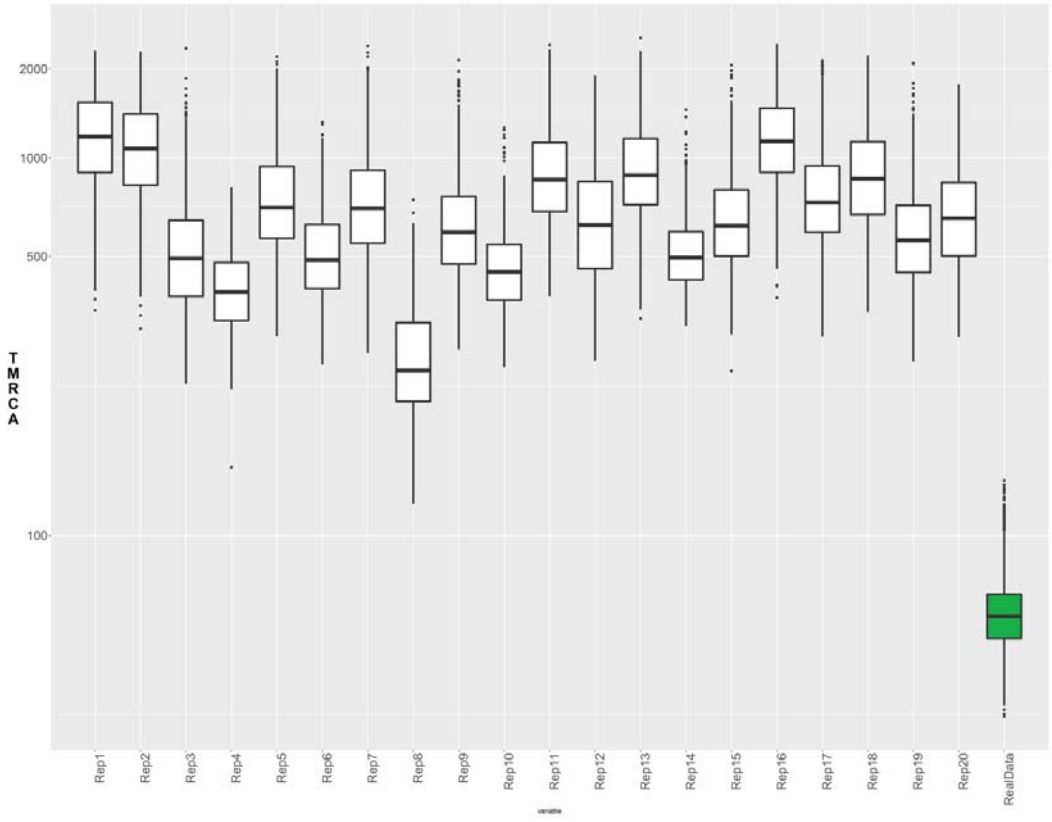


Figure S3. Calculated TMRCA of all isolates following tip-randomization.

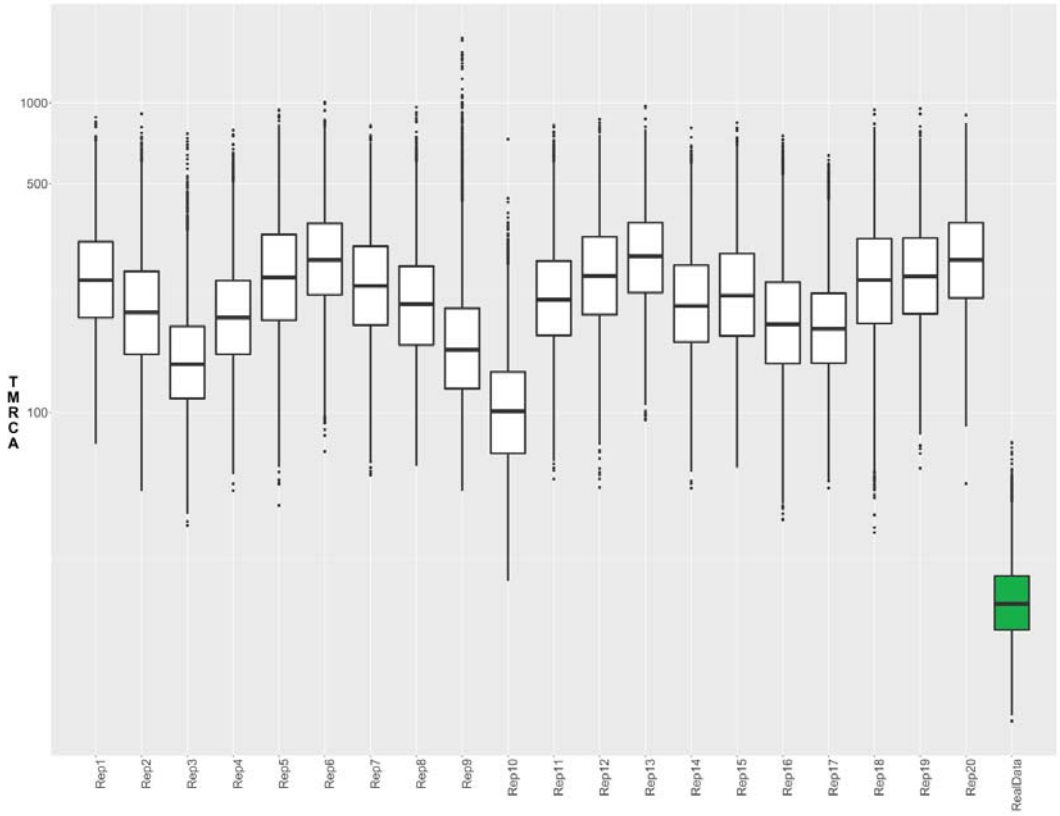


Figure S4. Calculated TMRCA of the Afghan strain family following tip-randomization.

Figure S5. Tipdate-calibrated Beast phylogeny including all 85 isolates showing posterior probabilities of individual nodes

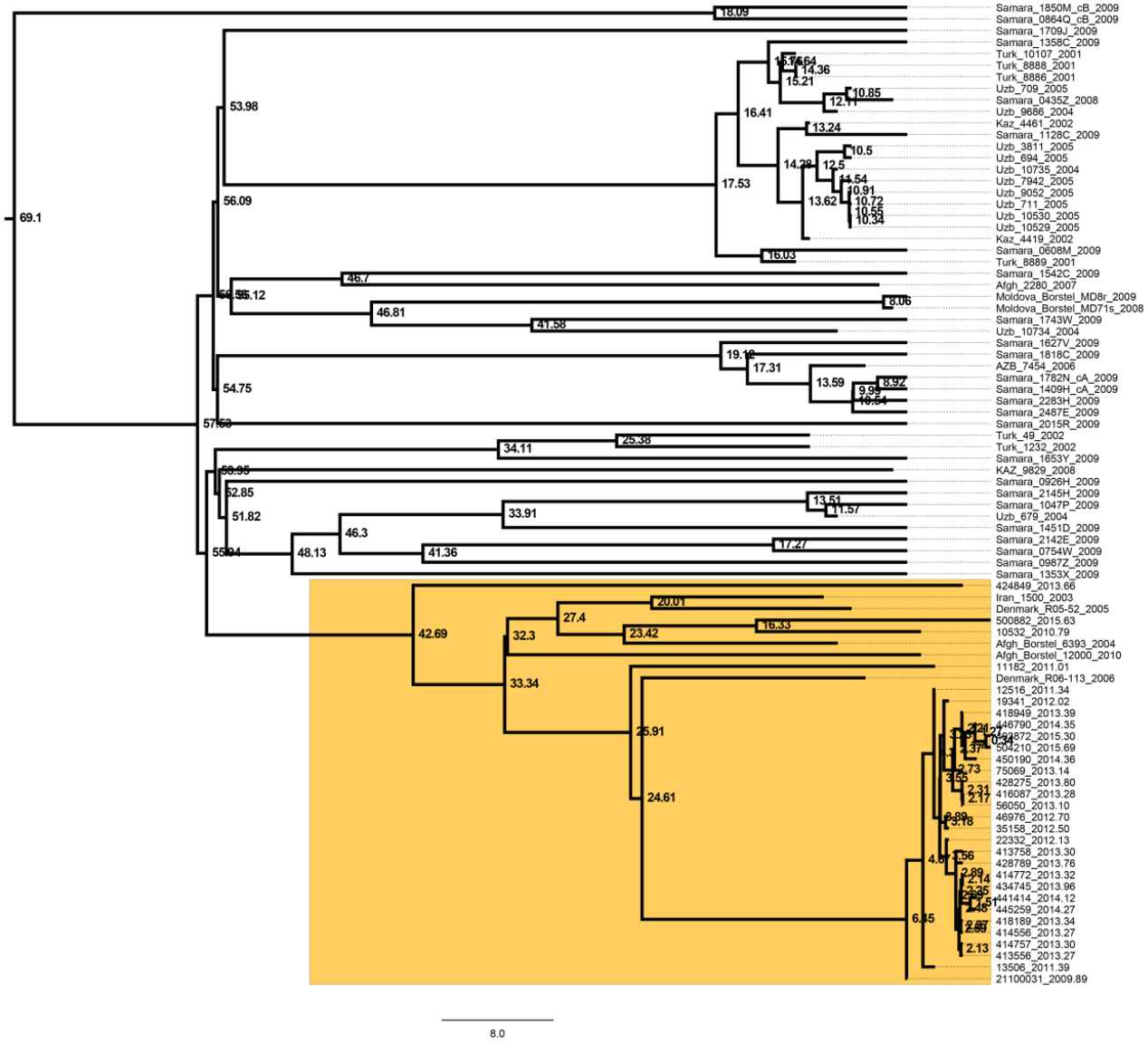


Figure S6. Tipdate-calibrated Beast phylogeny including all 85 isolates showing individual node ages.

Supplementary table S1. Model comparison using posterior simulation-based analog og Akaike's information criterion (AICM)

Afghan strain family

Demographic model comparison

	AICM	S.E.	Constant	Exponential	Logistic	Skyride	Expansion
Constant	32398179.2	+/- 0.133	-	-14.315	1.663	-28.481	-8.517
Exponential	32398164.9	+/- 0.148	14.315	-	15.978	-14.166	5.798
Logistic	32398180.9	+/- 0.154	-1.663	-15.978	-	-30.144	-10.18
Skyride	32398150.7	+/- 0.111	28.481	14.166	30.144	-	19.964
Expansion	32398170.7	+/- 0.128	8.517	-5.798	10.18	-19.964	-

Clock model comparison

			Strict	Lognorm relaxed
Strict	32398150.7	+/- 0.077	-	5.61
Lognorm relaxed	32398156.3	+/- 0.039	-5.61	-

Central Asian Clade

Clock model comparison

			Strict	Lognorm relaxed
Strict	32433074.8	+/- 0.165	-	-32.735
Lognorm relaxed	32433042.1	+/- 0.257	32.735	-