The molecular clock of *Mycobacterium tuberculosis*

F. Menardo[1,2*], S. Duchêne[3], D. Brites[1,2] and S. Gagneux[1,2]

[1] Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, 4002 Basel, Switzerland

[2] University of Basel, 4003 Basel, Switzerland

[3]Department of Biochemistry and Molecular Biology, Bio21 Molecular Sciences and Biotechnology Institute, University of Melbourne, 3010, Melbourne, Australia

* Correspondence: fabrizio.menardo@swisstph.ch

Abstract

The molecular clock and its phylogenetic applications to genomic data have changed how we study and understand one of the major human pathogens, *Mycobacterium tuberculosis* (MTB), the causal agent of tuberculosis. Genome sequences of MTB strains sampled at different times are increasingly used to infer when a particular outbreak begun, when a drug resistant clone appeared and expanded, or when a strain was introduced into a specific region. Despite the growing importance of the molecular clock in tuberculosis research, there is a lack of consensus as to whether MTB displays a clocklike behavior and about its rate of evolution. Here we performed a systematic study of the MTB molecular clock on a large genomic data set (6,285 strains), covering most of the global MTB diversity and representing different epidemiological settings. We found wide variation in the degree of clocklike structure among data sets, indicating that sampling times are sometimes insufficient to calibrate the clock of MTB. For data sets with temporal structure, we found that MTB genomes accumulate between $1x10^{-8}$ and $5x10^{-7}$ nucleotide changes per-site-per-year, which corresponds to 0.04 - 2.2 SNPs per-genome-per-year. Contrary to what expected, these estimates were not dependent on the time of the calibration points as they did not change significantly when we used epidemiological isolates (sampled in the last 40 years) or ancient DNA samples (about 1,000 years old) to calibrate the tree. Additionally, the uncertainty and the discrepancies in the results of different methods were often large, highlighting the importance of using different methods, and of considering carefully their assumptions and limitations.

Keywords

Evolution, Phylogenetics, Pathogen, Bacteria

Significance Statement

One of the major recent advancement in evolutionary biology is the development of statistical methods to infer the past evolutionary history of species and populations with genomic data. In the last five years, many researchers have used the molecular clock to study the evolution of *Mycobacterium tuberculosis*, a bacterial pathogen that causes tuberculosis and is responsible for millions of human deaths every year. The application of the molecular clock to tuberculosis is extremely useful to understand the evolution of drug resistance, the spread of different strains and the origin of the disease. Since some of these studies found contrasting results, we performed a systematic analysis of the molecular clock of MTB. This study will provide an important guideline for future analyses of tuberculosis and other organisms.

70 # Introduction

71 In 1962, Zuckerland and Pauling used the number of amino-acid differences among hemoglobin
72 sequences to infer the divergence time between human and gorilla, in what was the first application of
73 the molecular clock (Zuckerland and Pauling 1962). Although many at the time found it "crazy"
74 (Morgan 1998), soon the molecular clock was incorporated in Kimura's neutral theory of molecular
75 evolution (Kimura 1968), and found its place in the foundations of evolutionary biology. Thanks to the
76 improvements of sequencing technologies and statistical techniques it is now possible to use sequences
77 sampled at different times to calibrate the molecular clock and study the temporal dimension of
78 evolutionary processes in so called measurably evolving populations (Drummond et al. 2003). These
79 advancements have been most relevant for ancient DNA (aDNA), and to study the evolutionary
80 dynamics of pathogen populations, including one of the deadliest human pathogens: *Mycobacterium*
81 *tuberculosis* (WHO 2018).

82 In 1994, Kapur and colleagues pioneered molecular clock analyses in MTB: they assumed a clock rate
83 derived from other bacteria and used genetic polymorphisms to infer the age of divergence of different
84 MTB strains (Kapur et al. 1994). Since then, phylogenetic analyses with a molecular clock have been
85 used to estimate the timing of the introduction of MTB clades to particular geographic regions, the
86 divergence time of the MTB lineages, and the age of the most recent common ancestor (MRCA) of the
87 MTB complex (MTBC)(Comas et al. 2013, Bos et al. 2014, Merker et al. 2015, Kay et al. 2015,
88 Brynildsrud et al. 2018, Liu et al. 2018, Rutaihwa et al. 2018, O'Neill et al. 2018). Clock models,
89 together with phylodynamic models in a Bayesian setting have been used to characterize tuberculosis
90 epidemics by determining the time in which outbreaks began and ended (Eldholm et al. 2015, Lee et al.
91 2015, Folkvardsen et al. 2017, Bainomugisa et al. 2018), establishing the time of origin and spread of
92 drug resistant clades (Cohen et al. 2015, Eldholm et al. 2015, Eldholm et al. 2016, Brynildsrud et al.
93 2018), and correlating population dynamics with historical events (Pepperell et al. 2013, Merker et al.

94 2015, Eldholm et al. 2016, Liu et al. 2018, O'Neill et al. 2018, Merker et al. 2018). One example of the
95 potential of molecular clock analyses is the study of Eldholm and colleagues (Eldholm et al. 2016),
96 where the collapse of the Soviet Union and of its health system was linked to the increased emergence
97 of drug resistant strains in former Soviet countries, thus providing an insight on the evolutionary
98 processes promoting drug resistance, and informing control strategies.

99 A key aspect about estimating evolutionary rates and timescales in microbial pathogens is assessing
100 their clocklike structure. All molecular clock analyses require some form of calibration. In many
101 organisms this consists in constraining internal nodes of phylogenetic trees to known divergence times
102 (for example, assuming codivergence between hosts, or the fossil record), but in rapidly evolving
103 pathogens and studies involving ancient DNA, it is also possible to use sampling times for calibrations
104 (Seo et al. 2002). In the latter approach, the ages of tips of the tree, rather than those of internal nodes
105 are constrained to their collection times. Clearly, the sampling time should capture a sufficient number
106 of substitutions to estimate the evolutionary rate, which will depend on the evolutionary rate of the
107 organism and the extent of rate variation among lineages. Some popular methods to assess such
108 clocklike structure are the root-to-tip regression and the date randomization test (DRT).

109 While many of the studies inferring evolutionary rates for MTB reported support for a molecular clock
110 (Eldholm et al. 2015, Kay et al. 2015, Eldholm et al. 2016, Folkvardsen et al. 2017, Brynildsrud et al.
111 2018, Merker et al. 2018, Rutaihwa et al. 2018), some found a lack of clocklike structure (Comas et al.
112 2013, Bainomugisa 2018), and others assumed a molecular clock without testing whether the data had a
113 temporal structure (Pepperell et al. 2013, Cohen et al. 2015, Merker et al. 2015, Lee et al. 2015, Liu et
114 al. 2018, O'Neill et al. 2018). Based on a meta-analysis, Eldholm and colleagues (Eldholm et al. 2016)
115 observed that, excluding Comas et al. 2013, where the clock was calibrated assuming co-divergence
116 between MTB lineages and human mitochondrial haplotypes (i.e. internal node calibrations), and was
117 estimated to be around 10^{-9} nucleotide changes per site years, in all other studies, where the calibration

118 was based on the sampling time (tip-dating), the clock rate estimates spanned roughly an order of
119 magnitude around 10^{-7} nucleotide changes per site per year. They also showed that lineage 2 (L2) had a
120 faster clock rate compared to lineage 4 (L4), and that studies based on aDNA produced lower clock rate
121 estimates (Bos et al. 2014, Kay et al. 2015), thus suggesting a time dependency of clock rates in MTB.
122 The analysis of Eldholm and colleagues (Eldholm et al. 2016) highlighted the importance and the
123 difficulties of comparing results of molecular clock analyses from different data sets, it has however
124 two main limitations: 1) the comparison included mostly L4 data sets, only one L2 data set, and none
125 from other lineages, and 2) Eldholm et al. (2016) performed a meta-analysis (i.e the clock rates were
126 taken directly from the published studies), therefore the observed differences could be due to different
127 rates of molecular evolution among MTB populations, to methodological discrepancies among studies,
128 or a combination of both.

129 Here, we assembled a large genomic data set including sequences from all major lineages of MTB
130 (6,285 strains in total, belonging to six human adapted lineages, L1-L6, and one lineage predominantly
131 infecting cattle, *M. bovis*). We then applied the same set of methodologies to the whole data set, to
132 individual lineages and sub-lineages, and to selected local outbreaks, thus ensuring the comparability of
133 the results among different clades and epidemiological settings.

134 With this systematic approach, we addressed the following questions:

135 1) Is there a molecular clock in MTB and how do we detect it?

136 2) What is the clock rate of MTB, and what is its variation among lineages, sub-lineages and individual
137 outbreaks?

138 3) Are clock rate estimates dependent on the age of the calibration points in MTB?

139 4) What is the best approach for molecular clock analyses in MTB?

Results and Discussion

Is there a molecular clock in MTB?

Finding evidence of temporal structure is the first step when performing molecular clock analyses (Rieux and Balloux 2016). If there is not enough genetic variation between samples collected at different times, these cannot be used to calibrate the molecular clock, i.e. the population is not measurably evolving. To test temporal structure of MTB data sets we identified 6,285 strains for which the date of isolation was known, and with a good quality genome sequence (Methods, Table S1).

Temporal structure of the MTB lineages

We used root to tip regression to evaluate the temporal structure of the whole MTB complex and of the individual lineages (L1-L6 and *M. bovis*) (Rambaut et al. 2016). The root to tip regression is a regression of the root-to-tip distances as a function of sampling times of phylogenetic trees with branch lengths in units of substitutions per site, where the slope corresponds to the rate. Under perfect clock-like behavior, the distance between the root of the phylogenetic tree and the tips is a linear function of the tip's sampling year: recently sampled strains are further away from the root than older ones, such that the R^2 is the degree of clocklike behavior (Korber et al. 2000). We obtained very low values of R^2 for all lineages (maximum 0.1 for *M. bovis*), indicating a lack of clock-like behavior (Fig. S1). Additionally, we found a weak negative slope for L1, L5 and L6, normally interpreted as evidence for a lack of temporal structure, or overdispersion in the lineage-specific clock rates (Rambaut et al. 2016, Fig. S1, Table S2).

Since root to tip regression can be used only for exploratory analyses and not for formal hypothesis testing (Rambaut et al. 2016), we performed a DRT. The DRT consists in repeatedly reshuffling the date of sampling among taxa and then comparing the clock rate estimates among the observed and

163 reshuffled data sets (Ramsden et al. 2008). If the estimation obtained from the observed data does not
164 overlap with the estimations obtained from the randomized data sets, we can conclude that the observed
165 data has a stronger temporal signal than expected by chance, such that there is statistically significant
166 clocklike structure (Ramsden et al. 2008). Usually the DRT is implemented in a Bayesian phylogenetic
167 setting, however, considered the size and the number of data sets included in this study, an excessive
168 amount of computation would be required. To overcome this problem, we estimated the clock rate with
169 the least-squared dating method implemented in LSD (To et al. 2015). The advantage of this method is
170 that it is fast, therefore it can be used on data sets with thousands of taxa, and with more
171 randomizations compared to the10-20 typically used in a Bayesian setting (Duchene et al. 2018). One
172 limitation of least squares dating is that it typically assumes a single tree topology and vector of branch
173 lengths, however, a simulation study showed that maximum likelihood trees produced similar estimates
174 compared to the true topology (To et al. 2015). An additional limitation of LSD is that it assumes a
175 strict clock (i.e. all branches have the same clock rate), however it was shown to be robust to
176 uncorrelated variation of the clock rate among branches in the phylogeny (To et al. 2015). For each
177 data set we reshuffled the year of sampling among tips 100 times and estimated the clock rate of
178 observed and randomized data sets with LSD. All eight data sets except L5 and L6 passed the DRT
179 (Methods, Figs. S1-S5, Table S2). L5 and L6 are the two lineages with the lowest sample size, 117 and
180 33 strains respectively, moreover most strains were sampled in a short temporal period compared to the
181 other lineages (Figs. S6-S7). It is likely that with additional strains sampled across a larger time period,
182 L5 and L6 will also show evidence for a molecular clock. It was more surprising to find that L1 passed
183 the DRT despite having a negative rate in the root to tip regression. This pattern, however, is not
184 uncommon because the root-to-tip regression is not constrained on the phylogenetic tree, such that it
185 sometimes fails to locate an appropriate position for the root (Tong et al. 2018), particularly if there is a

186 significant amount of rate variation among sub-lineages of L1. These results suggest that a DRT is a
187 valuable test, even for data sets that show a negative slope value in the root to tip regression.

188 We complemented the analysis described above with a Bayesian phylogenetic analysis in Beast2
189 (Bouckaert et al. 2014). Since this is computationally very intense, we reduced the largest data sets
190 (MTBC, L1, L2 and L4) to 500 strains using two sub-sampling strategies: 1) we selected 500 random
191 strains 2) we used Treemmer to select 500 strains representative of the phylogenetic diversity (Menardo
192 et al. 2018). In most cases, the two strategies produced comparable results (Figs. S8-S9), here we report
193 the results for the randomly selected subsets (see SI Appendix for a comparison of the two strategies).
194 We assumed a coalescent constant population size as tree prior, a relaxed clock model, and a uniform
195 prior for the clock rate, constrained between 10^{-10} and 10^{-5} nucleotide changes per site per year. Such
196 interval spans the range of clock rates proposed for *M. tuberculosis* and for most other bacteria
197 (Duchene et al 2016 b, Eldholm et al. 2016). We observed that for all data sets the posterior is clearly
198 distinct from the prior, thus indicating that the data is informative (Drummond et al. 2006). Again, the
199 only exceptions were L5 and L6, where the posterior distribution is less distinct from the prior, and it
200 almost reaches the lower bound of 10^{-10} nucleotide changes per site per year, confirming the low
201 temporal structure of these two data sets.

202 **Temporal structure in sub-lineages and outbreaks**

203 We repeated these analyses on 23 sub-lineages and 7 outbreaks and local populations to test whether
204 we could detect a temporal structure also in smaller, less diverse data sets. With this sub-sampling
205 scheme we could compare the results among different clades, among outbreaks with different
206 epidemiological characteristics, and among local outbreaks and global data-sets (see Methods). We
207 found that 11 sub-lineages and 5 local populations passed the DRT (Table S2, Figs. S2-S7 and S10-
208 S12). All the data sets that failed the DRT had less than 350 genomes, or were composed of strains

sampled in a temporal range of 20 years or less. Additionally, only six of the 23 data sets with less than 350 strains or sampled across 20 years or less, passed the DRT (Fig. 1).

Among the three methods generally used to study the temporal structure of a data set, the root to tip regression was the least reliable as it failed to detect the temporal structure of some of the data sets that passed the DRT, but had a negative slope of the regression (L1, L4.1.2, L1.1.1). Nevertheless, root to tip regression can be useful to identify data sets where the temporal signal comes from a single strain, or a few strains (SI Appendix). Comparing prior and posterior distributions of the clock rates was also useful to detect the presence of temporal structure, and in our case all data sets that passed the DRT had a posterior distribution of the clock rate clearly distinct from the prior. However, no formal test exists to compare posterior and prior, and visual inspection lacks consistency (Bromham et al. 2018). In our case, some of the data sets that did not pass the DRT (L2.2.1_nc1, L2.2.1_nc2, Trewby 2016) had a posterior distribution of the clock rate as distinct from the prior as some of the data set that passed the DRT (Figs. S2 and S4). It is possible that Beast has stronger statistical power and can detect temporal structure in data sets where the DRT with LSD fails to do so, however it was shown that when the tree prior is misspecified, the deviation of the posterior distribution of the clock rate from the prior can be an artifact and not the result of genuine temporal structure (Möller et al. 2018).

What is the clock rate of MTB, and what is its variation among lineages, sub-lineages and outbreaks?

We found that the point estimates of all data sets that passed the DRT are included in one order of magnitude, between 2.99×10^{-8} (L4.10 LSD) and 3.17×10^{-7} (L1.2.2 LSD) nucleotide changes per site per year. While the data sets with the weakest temporal structure had a low range of the 95% confidence interval (CI) reaching the hard limit imposed by LSD of 10^{-10}, most of the CI and 95% highest posterior

232 density (HPD) ranges are included between 10^{-8} and $5x10^{-7}$ (Fig. 2 and Table S2). This range
233 encompasses previous estimates obtained with epidemiological samples and aDNA (Eldholm et al.
234 2016) and is among the lowest in bacteria (Duchene et al. 2016b). Because of the slow clock rate, tip-
235 dating with MTB requires samples collected over long periods: most data sets with a temporal structure
236 included strains sampled over 20 years. Additionally, a large sample size can increase the statistical
237 power, for example the complete L2.2.1 data set, composed of 2,115 genomes sampled over 22 years
238 passed the DRT, while the trimmed sub-sample of 500 strains sampled over the same time period did
239 not pass it (Fig. S8 and Table S2).

240 There are two notable exceptions to the pattern described above: the data sets L4_nc and Eldholm
241 2016 which produced much higher clock rate estimates compared to all other data sets included in this
242 study (Table S2). However, there are several reasons to doubt their accuracy, especially because they
243 are the smallest among all considered data sets, with 32 and 33 strains respectively, and show a high
244 degree of temporal and phylogenetic clustering (SI Appendix). We excluded the results of L4_nc from
245 further analyses because they are most likely an artifact due to the low sample size. Conversely, we
246 kept the data set from Eldholm et al. (2016) for further analyses, although its estimates, and especially
247 the Beast results, should be taken with much caution (SI Appendix).

248 The large HPD and CI, and the diverging results between LSD and Beast of some data sets, make it
249 difficult to pinpoint specific differences between lineages, sub-lineages and local populations (Fig. 2).
250 For example, L1 and *M. bovis* data sets have point estimates higher than most L4 data sets, but the CI
251 and HPD are often overlapping. Moreover, when considering only the results of Beast, L2 had a higher
252 clock rate compared to L4, as suggested by Eldholm et al. (2016), and all data sets included in the sub-
253 lineage L2.2.1 showed a faster clock rate compared to the complete L2 data set (Fig. 2). The sub-
254 lineage L2.2.1 includes the so called "modern Beijing" family, which was shown to be
255 epidemiologically associated with increased transmission, virulence and drug resistance (Glynn et al.

256 2006, Hanekom et al. 2010, de Steenwinkel et al. 2012, Ribeiro et al. 2014, Holt et al., Wiens et al.
257 2018), and to have a higher mutation rate compared to L4 strains (Ford et al. 2013). However, the LSD
258 estimates for L2.2.1 and for its sub-lineages, despite showing the same trend of Beast, have large
259 confidence intervals, overlapping with the results of L2 and L4 (Fig. 2). In summary, Beast estimates
260 suggests that L2.2.1 has a faster clock rate compared to L4, possibly because of the higher mutation
261 rate, while LSD indicates that the clock rates of L4 and L2.2.1 are similar (Fig. 2).

262 Further evidence of among-lineage variation is provided by the results of the Bayesian analyses, where
263 for most data sets we obtained coefficient of variations (COVs) with a mode of 0.2 – 0.3, and not
264 abutting zero (Table S2), thus indicating a moderate deviation from a strict clock (Drummond et al.
265 2006). Taken together, these results indicate that there is a moderate variability among the current rate
266 of molecular evolution in MTB. However, the clock rate variation of data sets with different genetic
267 background (lineages), genetic diversity, geographic spread, infected host, epidemiological setting or
268 presence of drug resistance mutations is contained within one and a half order of magnitude (10^{-8} -
269 $5x10^{-7}$), and it could not be fully resolved with these data and inference methods.

270 A possible confusing factor in our results is the influence of the tree prior on the results of the Bayesian
271 analysis (Crispell et al. 2017). We already mentioned that the comparison of the prior and posterior
272 distributions of the clock rate showed that the posterior is distinct from the prior and therefore the
273 results of the Bayesian analyses are mostly informed by the data, and not by the prior (Bromham et al.
274 2018). An alternative way to investigate the sensitivity to the prior is to use different ones, and compare
275 the results. We did this with the five outbreaks that passed the DRT and found that in some cases the
276 different priors affected moderately the clock rate estimates (SI Appendix, Table S3), although this does
277 not change the general picture described above. We also found that the tree age estimates can diverge
278 drastically with different tree priors, and we will address this further below.

279

280 **A faster clock for the ancestor of *M. bovis*?**

281 Above we showed that current clock rates are relatively similar among different data sets; a different
282 question is whether the clock rate was constant during the evolutionary history of the MTB complex.
283 When looking at the phylogenetic tree of the MTB complex, rooted with the genome sequence of *M.*
284 *canettii,* one can notice that strains belonging to different lineages, despite being all sampled in the last
285 40 years have different distances from the root (Fig. 3). For example, since their divergence from the
286 MRCA of the MTB complex, the two *M. africanum* lineages (L5 and L6) and especially *M. bovis,*
287 accumulated more nucleotide changes than the lineages belonging to MTB sensu stricto (L1-L4)
288 (Figure 3). Additionally, all methods (root to tip regression, LSD and Beast) if used without an
289 outgroup, placed the root on the branch between *M. bovis* and all other lineages, while rooting the tree
290 with the outgroup *M. canettii* placed the root on the branch connecting MTB sensu stricto with *M.*
291 *africanum* (L5 and L6) and *M.bovis.* The different root placement affects the clock rate estimation only
292 moderately (4.16×10^{-8} LSD analysis without outgroup, 5.59×10^{-8} LSD analysis with outgroup, Table
293 S2), but it is a further indication of the variation of the rate of molecular evolution during the
294 evolutionary history of the MTB complex. The observation that all *M. bovis* strains, despite having a
295 clock rate similar to all other data sets, have a larger distance from the root of the MTB complex tree
296 compared to other lineages is intriguing, and could be explained by a faster rate of molecular evolution
297 of the ancestors of *M. bovis* (Figs. 2-3). It is believed that *M. bovis* switched host (from human to
298 cattle) (Brosh et a. 2002, Mostowy et al. 2002, Brites et al. 2018), and it is possible that during the
299 adaptation to the new host several genes were under positive selection, thus leading to an increase in
300 the accumulation of substitutions in the *M. bovis* genome. Another possibility is that the ancestor of *M.*
301 *bovis* experienced a period of reduced population size, a bottleneck, in such case the clock rate is

302 predicted to accelerate, under the increased influence of genetic drift that fixes mutation with at a faster
303 pace compared to populations with a larger population size.

304

305 Time dependency of the clock rate

306 It has been suggested that in MTB, as in other organisms, the clock rate estimation is dependent on the
307 age of the calibration points (Ho et al. 2005, Ho et al. 2011, Comas et al. 2013, Duchene et al. 2014,
308 Eldholm et al. 2016, Duchene et al. 2016 b), and that using recent population-based samples could
309 result in an overestimation of the clock rate, because these samples include deleterious mutations that
310 have not yet been purged by purifying selection. We tested this hypothesis by repeating the analyses
311 presented above, only this time we included the aDNA genome sequences of three MTB strains
312 obtained from Precolumbian Peruvian human remains (Bos et al. 2014). If the clock rate estimates
313 depend on the age of the calibration points, adding ancient genomes should results in lower clock rates.
314 We performed this analysis with LSD, using the complete data set (6,285 strains), and with Beast, using
315 the sub-sample of 500 randomly selected strains. With both methods we found that the clock rate
316 estimation for the data sets with the aDNA samples are overlapping with the estimates obtained without
317 the aDNA samples (Fig. 4, Table S2). These results indicate that there is no significant effect of the age
318 of the calibration points on the clock rate estimations of MTB (at least in these data), and are in contrast
319 with the finding of previous studies (Comas et al. 2013, Duchene et al. 2016b, Eldholm et al. 2016).
320 However, those previous studies compared the results obtained from different data sets of MTB and
321 other bacteria, and did not investigate the direct impact of adding older calibration points to a data set.
322 It is possible that the effect of the dependency of the rate estimates from the age of calibration is small,
323 and our analysis lacked the statistical power to detect it. This could be due to weak purifying selection
324 in MTB caused by the lack of recombination (Hershberg et al. 2008). Under this hypothesis, most of

325 the slightly deleterious mutations accumulated in the last 1,000 years have not been purged by
326 purifying selection. Some support to this hypothesis is provided by the observation that MTB mutation
327 rates in vitro and vivo, estimated with fluctuation assays and resequencing of strains infecting
328 macaques, are remarkably similar to the clock rates obtained in this study (~ $3x10^{-8}$ - $4x10^{-7}$)(Ford et al.
329 2011).

330 The aDNA samples considered in this study are not optimal to test the time dependency hypothesis
331 because they belong to the *M. pinnipedii* (Bos et al. 2014). The modern strains of this lineage are
332 rarely sampled, because they are infecting seals rather than humans. Additional aDNA samples from
333 older periods and belonging to other lineages are necessary to better investigate the time dependency
334 hypothesis in MTB.

335

336 Dating MTB phylogenies

337 In most cases, the goal of molecular clock studies is not to estimate the clock rates, but rather the age of
338 the phylogenetic tree and of its nodes. Conceptually, this means extrapolating the age of past events
339 from the temporal information contained in the sample set. If we exclude the few aDNA samples that
340 are available (Bos et al. 2014, Kay et al. 2015), all MTB data sets have been sampled in the last 40
341 years. It is therefore evident that the age estimates of recent shallow nodes will be more accurate than
342 medium and deep nodes, and in part this is reflected in the larger CI and HPD of the age of ancient
343 nodes compared to more recent ones. Extrapolating the age of trees that are thousands of years old with
344 contemporary samples is particularly challenging, because the observed data captures only a small
345 fraction of the sample's evolutionary history, and these are the cases where aDNA samples are most
346 valuable. After these general words of caution, we will discuss here the implications of our results on

347 the dating of different MTB trees. At first, we consider tip dating of MTB outbreaks and local
348 populations, where the age of the tree is relatively recent, and in principle it can be accurately inferred.

349 **Dating MTB outbreaks and local populations**

350 For the five local data sets included in this study that passed the DRT, the results of Beast and LSD
351 were quite similar, even when the two methods produced different clock rates, like for Eldholm 2016
352 and Folkvardsen 2017 (Fig. 2 and S15, Table S2). The MRCA of the New Zealand population of
353 *M.bovis* (Crispell 2017) was estimated to be approximately 400 to 800 years old, and the MRCAs of
354 the outbreaks (Lee 2015, Eldholm 2015, Eldholm 2016, Folkvardsen 2017) were estimated to be at
355 most 200 years old, the only exception being the LSD results of Eldholm 2016 which have extremely
356 broad CI spanning almost 250,000 years, probably due to the small sample size (Table S2, Fig. S15).
357 However, when we considered the effect of the tree prior on the results of the Bayesian analysis, we
358 found that both LSD and Beast (used with a coalescent constant population size prior) overestimated
359 the age of the MRCA of data sets which showed evidence for population growth (Eldholm 2015,
360 Crispell 2017 and Folkvardsen 2017), compared to Beast used with and exponential population growth
361 prior (SI Appendix, Table S3). The most striking case was the Argentinian outbreak (Eldholm 2015),
362 which showed evidence for strong population growth, and for which the analyses with a coalescent
363 exponential growth prior resulted in an age of the MRCA three times younger than the one obtained
364 with a coalescent constant population size prior (SI Appendix, Table S3). Through the analysis of the
365 posterior distribution of the growth rate, we showed that the exponential growth prior fits better
366 compared to the constant size, and this is supported by the Bayesian model test performed by Eldholm
367 and colleagues in the original publication (Eldholm et al. 2015). Beast (with an exponential growth
368 prior) estimated the MRCA of the Argentinian outbreak to be included between 1963 and 1986, and
369 LSD estimated it to be between 1887 and 1960. We cannot formally test which of the two estimates is
370 more accurate. However, the results of Beast showed that there is a high intra-lineage variation of the

371 clock rate (median COV = 0.83), and since LSD assumes a strict clock, in this case we cautiously
372 propend for the Beast results. This example shows how a careful consideration of different
373 phylogenetic methods, models and their assumptions, is fundamental to obtain reliable estimates of the
374 age of the MRCA of outbreaks and local population, and how failing to do so can have drastic
375 consequences.

376 **Dating MTB lineages**

377 A more delicate and complex question is the age of the MRCA of the MTB complex and of its lineages,
378 this topic is hotly debated, and highly relevant to understand the emergence and evolution of this
379 pathogen (Wirth et al. 2008, Comas et al. 2013, Bos et al. 2014). The LSD analyses that rooted the tree
380 with *M. canettii* estimated the MRCA of the MTB complex to be between 2,828 and 5,758 years old
381 (Table S3). These estimates are highly similar to the ones of Bos and colleagues (2,951 – 5,339) which
382 were obtained with Bayesian phylogenetics and a much smaller sample size (Bos et al. 2014).
383 Nevertheless, they should be taken with much caution because of the intrinsic uncertainty in estimating
384 the age of a tree that is several thousands of years old, calibrating the molecular clock with the
385 sampling time of modern strains and only 3 aDNA samples. A more approachable question is the age of
386 the MRCA of the individual MTB lineages. Here we can consider the results of four different analyses:
387 the LSD and Beast analyses on the individual lineages (L1-L4, and *M. bovis*), and the LSD and Beast
388 analyses on the complete MTB complex (including the aDNA samples), from which the age of the
389 MRCA of the lineages can be extracted (L1-L6, and *M. bovis*). When we combined all these results,
390 merging the CI and HPD, we obtained an estimate of the age of the MTB lineages which accounts for
391 the uncertainty intrinsic in each analysis, but also for the differences among inference methods and
392 models, thus providing a more conservative hypothesis. In all our analyses, the age of all lineages
393 resulted to be 4,000 years old or less, the only exception was the age of the MRCA of L2 for which the
394 95% CI of the LSD analysis extended to about 11,000 years ago (Fig. 5, Table S4). The large CI of L2

395 was maybe due to among-lineage variation of the clock rate in L2, as discussed above. While L5, L6
396 and *M. bovis* have younger MRCA and narrower confidence intervals, we should note that for these
397 lineages the sampling is much less complete compared to L1-L4, and it is possible that further
398 sampling will add more basal strains to the tree, thus resulting in older MRCA for these lineages. For
399 the other lineages, where the sampling is more representative of the global diversity, the confidence
400 interval of the age of the MRCA extends over several thousands of years, about 10,500 for L2, 3,500
401 for L1 and L3 and 3,000 for L4, and the point estimates of the four analyses spread over more than
402 1,000 years. These results suggest that we should be very careful when interpreting the results of tip
403 dating in MTB (and probably in other organisms as well), especially when using a single inference
404 method, and we should not be over confident on the accuracy of the obtained estimates.

405 Altogether our results highlighted the uncertainty of tip-dating, they nevertheless support the results of
406 Bos et al. 2014 that found the MRCA of the MTB complex to be relatively recent, and not compatible
407 with the out of-Africa-hypothesis (Wirth et al. 2008, Comas et al. 2013) in which the MTB lineage
408 differentiated in concomitance with the dispersal of *Homo sapiens* out of Africa, about 70,000 years
409 ago.

410 Dating analyses based on DNA samples can only reconstruct the evolutionary history of the data set as
411 far back as the MRCA of the sample. It is possible that in the future new lineages will be sampled, and
412 the MTB phylogeny will be updated moving the MRCA further in the past. Additionally, it is also
413 possible that extinct lineages were circulating and causing diseases much earlier that the MRCA of the
414 strains that are circulating now. This hypothesis is supported by the detection of molecular markers
415 specific for MTB in archeological samples (reviewed in Brites and Gagneux 2015), the oldest of them
416 in a bison's bone about 17,500 years old (Rothschild et al. 2001). Several such studies directly
417 challenge the results of tip-dating presented here, because they reported molecular marker specific to
418 MTB lineages in archeological samples that predate the appearance of those lineages as estimated by

419 tip dating (Taylor et al. 2007, Hershkovitz et al. 2008, Nicklisch et al. 2012). However, there is a
420 controversy regarding the specificity of some of the used markers, and the potential contamination of
421 some of the samples by environmental mycobacteria (Wilbur et al. 2009, Donoghue et al. 2009).

422 Whole genome sequences from additional aDNA samples are needed to reconcile these two diverging
423 lines of evidence, ideally they should belong to different lineages, represent different periods, and
424 include samples much older than the currently available aDNA from Peruvian human remains.

425

426 Which method to use?

427 The large differences between some estimates of LSD and Beast are a key result of this study.
428 Although we did not find any systematic bias (SI Appendix, Figs. S13-S15 , Table S5), and for most
429 data sets the two methods produced similar results, in some cases these differences can lead to
430 drastically different interpretations. For example, let us consider the age of the MRCA of L3: LSD
431 estimated it to be between 1,067 and 631 years old, while based on Beast it was between 4,081 and
432 1,060 years old. The two intervals are slightly overlapping, but if we gave in to the temptation of
433 relating these results with events in human history, as it is often done, we would depict very different
434 scenarios. Unfortunately, there is not an easy solution to this problem. Both methods have their pitfalls:
435 LSD does not consider uncertainty in the tree topology and it uses a uniform clock (To et al. 2015), on
436 the other end, Beast has more realistic clock models, and it can perform Bayesian model test, but it was
437 shown that it can overestimate the clock rate if the tree prior is misspecified (Möller et al. 2018). Tree
438 prior misspecification could potentially affect many MTB data sets, especially the more diverse (i.e
439 lineages and sublineages), where the genealogy of the sample is the result of complex population
440 dynamics, population subdivisions and sampling biases, and none of the available tree prior is likely to
441 capture the tree-generating process. For such complex data sets, a uniform prior on the tree topology

442 and vector of branch lengths, where all trees are equally likely a priori, could be an alternative worth
443 exploring. An additional complication is that large data sets are difficult to analyze with Beast, and we
444 showed that different sub-sampling strategies can lead to different results (SI Appendix). Conservative
445 researchers might want to use both methods, if the size of the data set allows it, several model and prior
446 combinations should be tested in Beast (Bromham et al. 2018), and the final results of the two methods
447 can be combined in one range providing an estimation of the uncertainty of the clock rate or of the age
448 of some specific node of the tree.

449

450 # Methods

451 **Bioinformatic pipeline**

452 We identified 21,734 MTB genome sequences from the sequence read archive (Table S6). All genome
453 sequences were processed similarly to what described in Menardo et al. (2018).

454 We removed Illumina adaptors and trimmed low quality reads with Trimmomatic v 0.33
455 (SLIDINGWINDOW:5:20) (Bolger et al. 2014). We excluded all reads shorter than 20 bp and merged
456 overlapping paired-end reads with SeqPrep (overlap size = 15) (https://github.com/jstjohn/SeqPrep).

457 We mapped the resulting reads to the reconstructed ancestral sequence of the MTBC (Comas et al.
458 2013) using the mem algorithm implemented in BWA v 0.7.13 (Li and Durbin 2009). Duplicated reads
459 were marked by the MarkDuplicates module of Picard v 2.9.1
460 (https://github.com/broadinstitute/picard). We performed local realignment around Indel with the
461 RealignerTargetCreator and IndelRealigner modules of GATK v 3.4.0 (McKennaet al. 2010). We used
462 Pysam v 0.9.0 (https://github.com/pysam-developers/pysam) to exclude reads with alignment score
463 lower than (0.93*read_length)-(read_length*4*0.07)): this corresponds to more than 7 miss-matches

464 per 100 bp. We called SNPs with Samtools v 1.2 mpileup (Li 2008) and VarScan v 2.4.1 (Koboldt et
465 al. 2012) using the following thresholds: minimum mapping quality of 20; minimum base quality at a
466 position of 20; minimum read depth at a position of 7X; minimum percentage of reads supporting the
467 call 90%; no more than 90%, or less than 10% of reads supporting a call in the same orientation (strand
468 bias filter). SNPs in previously defined repetitive regions were excluded (Comas et al. 2013). We
469 excluded all strains with average coverage < 15 X. Additionally, we excluded genomes with more than
470 50% of the SNPs excluded due to the strand bias filter. Furthermore, we excluded genomes with more
471 than 50% of SNPs with a percentage of reads supporting the call included between 10% and 90%. We
472 filtered out genomes with phylogenetic SNPs belonging to different lineages or sub-lineages (only for
473 L4) of MTB, as this is an indication that a mix of strains could have been sequenced. To do this, we
474 used the diagnostic SNPs obtained from Steiner et al. 2014 and Stucki et al. 2016 for L4 sub-lineages.
475 We excluded all strains for which we could not find the date of isolation 1) in the SRA meta-
476 information, 2) in the associated publications, 3) from the authors of the original study after inquiry. We
477 divided all remaining strain by lineage (L1 -L6 and *M. bovis*), and excluded strains with a number of
478 called SNPs deviating more than three standard deviations from the mean of the respective lineage. We
479 built SNPs alignments for all lineages including only variable positions with less than 10% of missing
480 data. Finally, we excluded all genomes with more than 10% of missing data in the alignment of the
481 respective lineage. After all filtering steps, we were able to retrieve 6,285 strains with high quality
482 genome sequences and an associated date of sampling (Table S1).

483

484 **Dataset subdivision**

485 To perform a systematic analysis of the molecular clock in MTB we considered different data sets:

486 1) the complete data sets all (6,285 strains)

487 2) the different lineages of MTB (L1, L2, L3, L4, L5, L6, *M. bovis*)

488 3) the sub-lineages of L1 (L1.1.1, L1.1.1.1, L1.1.2, L1.1.3, L1.2.1 and L1.2.2) and L2 (L2.1, L2.2.1,
489 L2.2.2 and L2.2.1.1) as defined by Coll et al. 2014; the sub-lineages of L4 (L4.1.1, L4.1.2,
490 L4.1.3, L4.4, L4.5, L4.6.1 and L4.10) as defined by Stucki et al . 2016. Additionally, we identified two
491 L4 clades that were not classified by the diagnostic SNPs of Stucki et al. 2016 (L4_nc and L4.1_nc,
492 respectively, included into L4.6.2 and L4.1.2 as defined by Coll et al. 2014), and four sub-clades of
493 L2.2.1 that were not previously designated as sub-lineages (L2.2.1_nc1, L2.2.1_nc2, L2.2.1_nc3 and
494 L2.2.1_nc4) (Supplementary Figs. 10-12).

495 4) Selected data sets representing outbreaks or local populations that have been used for molecular
496 clock analyses in other studies

497 Lee et al. 2015 - Mj clade outbreak among a Inuit population in Canada (L4)

498 Eldholm et al. 2015 - Multi-drug resistant outbreak in Argentina (L4)

499 Eldholm et al. 2016 – Afghan family outbreak in Oslo (L2)

500 Trewby et al. 2016 – *M. bovis* in Northern Ireland

501 Crispell et al. 2017 – *M. bovis* in New Zealand

502 Folkvardsen et al. 2017 - C2/1112-15 outbreak in Denmark (L4)

503 Bainomugisa et al. 2018 – Multi-drug resistant outbreak on Daru island in PNG (L2)

504

505 **Evaluation of the temporal structure and LSD analysis**

506 For all data sets, we assembled SNPs alignments including variable positions with less than 10% of
507 missing data. We inferred phylogenetic trees with raxml 8.2.11 using a GTR model (-m GTRCAT -V

508 options). Since the alignments contained only variable positions, we rescaled the branch lengths of the
509 trees rescaled_branch_length = ((branch_length * alignment_lengths) / (alignment_length +
510 invariant_sites)), Duchene and colleagues (Duchene et al. 2018) showed that this method produced
511 similar results compared to ascertainment bias correction. We then used the R package ape (Paradis et
512 al. 2018) to perform root to tip regression after rooting the trees in the position that minimizes the sum
513 of the squared residuals from the regression line. Root to tip regression is only recommended for
514 exploratory analyses of the temporal structure of a dataset and it should not be used for hypothesis
515 testing (Rambaut et al. 2016). A more rigorous approach is the date randomization test (DRT)(Ramsden
516 et al. 2008), in which the sampling dates are reshuffled randomly among the taxa and the estimated
517 molecular clock rates estimated from the observed data is compared with the estimates obtained with
518 the reshuffled data sets. This test can show that the observed data has more temporal information that
519 data with random sampling times. For each dataset, we used the least square method implemented in
520 LSD v0.3-beta (To et al. 2015) to estimate the molecular clock in the observed data and in 100
521 randomized replicates. To do this, we used the QPD algorithm allowing it to estimate the position of the
522 root (option -r a) and calculating the confidence interval (options -f 100 and -s). We defined three
523 different significance levels for the DRT: 1) the simple test is passed when the clock rate estimate for
524 the observed data does not overlap with the range of estimates obtained from the randomized sets. 2)
525 The intermediate test is passed when the clock rate estimate for the observed data does not overlap with
526 the confidence intervals of the estimates obtained from the randomized sets. 3) The stringent test is
527 passed when the confidence interval of the clock rate estimate for the observed data does not overlap
528 with the confidence intervals of the estimates obtained from the randomized sets.

529

530 **Bayesian phylogenetic analysis**

531 Bayesian molecular clock analyses are computationally demanding and problematic to run on large
532 data sets. Therefore we reduced the seven largest data sets (MTBC, L1, L2, L2.2.1, L4, L4.1.2, L4.10)
533 to 500 representative strains in two different ways; 1) by randomly selecting 500 strains 2) using
534 Treemmer with -r 1 option (Menardo et al. 2018). We assembled SNPs alignments including only
535 variable positions with less than 10% of missing data and we performed Bayesian inference with
536 Beast2 (Bouckaert et al. 2014). We corrected the xml file to specify the number of invariant sites as
537 indicated here: https://groups.google.com/forum/#!topic/beast-users/QfBHMOqImFE, and used the tip
538 sampling year as calibration. We assumed the HKY model, the relaxed lognormal clock model
539 (Drummond et al. 2006) and the constant population size coalescent prior. We chose a lognormal(1,
540 1.5) prior for kappa, a 1/x prior for the population size [0- 10^9] a uniform prior for the mean rate [10^{-10}
541 – 10^{-5}] and a normal(0,10) prior for the standard deviation of the rate [0 – infinity]. For all data sets,
542 we ran at least two runs, we used Tracer 1.7.1 (Rambaut et al. 2018) to identify and exclude the burn-
543 in, to evaluate convergence among runs and to calculate the estimated sample size (ESS). We stopped
544 the runs when at least two chains reached convergence, and the ESS of the posterior and of all
545 parameters were larger than 200.

546

547 **Analyses with the complete MTB complex and aDNA**

548 We analyzed the complete data set of 6,285 genomes with the same methods described above. The only
549 difference was that for the LSD analysis, we rooted the input tree using *Mycobacterium canetti*
550 (SAMN00102920, SRR011186) as outgroup. We did this because we noticed that without outgroup, all
551 methods placed the root on the branch separating *M. bovis* from all other lineages, and not on the
552 branch separating MTB sensu stricto from the other lineages.

553 To test the time dependency hypothesis, we repeated the LSD and Beast analyses on the MTBC, adding
554 the aDNA genome sequences of three MTB strains obtained from Precolumbian Peruvian human
555 remains (Bos et al. 2014). These are the most ancient aDNA samples available for MTB. For LSD, we
556 assigned as sampling year the confidence interval of the radiocarbon dating reported in the original
557 publication. For Beast, we assigned uniform priors spanning the confidence interval but we failed to
558 reach convergence, therefore we used the mean of the maximum and minimum years in the confidence
559 interval (SAMN02727818: 1126 [1028-1224], SAMN02727820: 1117 [1023 - 1211] ,
560 SAMN02727821: 1211 [1141 – 1280]).

561 To summarize the results of the Beast analysis with the aDNA samples and retrieve the age of the
562 MRCA of the individual lineages, we randomly sampled 5,000 trees from the posterior (after excluding
563 the burn-in), and calculated the Maximum clade credibility tree with the software Treeannotator v2.5.0.

564

565 Acknowledgments

566 Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at
567 the University of Basel. This work was supported by the Swiss National Science Foundation (grants
568 310030_166687, IZRJZ3_164171, IZLSZ3_170834 and CRSII5_177163), the European Research
569 Council (309540-EVODRTB) and SystemsX.ch. SD was supported by a McKenzie fellowship from the
570 University of Melbourne.

571

572

573

References

Bainomugisa, A., Lavu, E., Hiashiri, S., Majumdar, S., Honjepari, A., Moke, R., ... & Coulter, C. (2018). Multi-clonal evolution of multi-drug-resistant/extensively drug-resistant Mycobacterium tuberculosis in a high-prevalence setting of Papua New Guinea for over three decades. *Microbial genomics*, *4*(2).

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, *30*(15), 2114-2120.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., ... & Campbell, T. J. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, *514*(7523), 494.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., ... & Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, *10*(4), e1003537.

Brites, D., Loiseau, C., Menardo, F., Borrell, S., Boniotti, M. B., Warren, R., ... & Fyfe, J. A. (2018). A New Phylogenetic Framework for the Animal-adapted Mycobacterium tuberculosis Complex. *Frontiers in Microbiology*, 9.

Bromham, L., Duchêne, S., Hua, X., Ritchie, A. M., Duchêne, D. A., & Ho, S. Y. (2018). Bayesian molecular dating: opening up the black box. *Biological Reviews*, *93*(2), 1165-1191.

Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., ... & Parsons, L. M. (2002). A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the national academy of Sciences*, *99*(6), 3684-3689.

Brynildsrud, O. B., Pepperell, C. S., Suffys, P., Grandjean, L., Monteserin, J., Debech, N., ... & Fandinho, F. (2018). Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. *Science Advances*, *4*(10), eaat5869.

Coll, F., McNerney, R., Guerra-Assuncao, J. A., Glynn, J. R., Perdigao, J., Viveiros, M., ... & Clark, T. G. (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nature communications*, *5*, 4812.

611 Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., ... & Gagneu, S. (2013).
612 Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern
613 humans. *Nature genetics*, *45*(10), 1176.
614

615 Crispell, J., Zadoks, R. N., Harris, S. R., Paterson, B., Collins, D. M., de-Lisle, G. W., ... & Kao, R. R.
616 (2017). Using whole genome sequencing to investigate transmission in a multi-host system: bovine
617 tuberculosis in New Zealand. *BMC genomics*, *18*(1), 180.
618

619 de Steenwinkel, J. E., Marian, T., de Knegt, G. J., Kremer, K., Aarnoutse, R. E., Boeree, M. J., ... &
620 Bakker-Woudenberg, I. A. (2012). Drug susceptibility of Mycobacterium tuberculosis Beijing
621 genotype and association with MDR TB. *Emerging infectious diseases*, *18*(4), 660.
622

623 Donoghue, H. D., Hershkovitz, I., Minnikin, D. E., Besra, G. S., Lee, O. Y. C., Galili, E., ... & Bar-Gal,
624 G. K. (2009). Biomolecular archaeology of ancient tuberculosis: response to "Deficiencies and
625 challenges in the study of ancient tuberculosis DNA" by Wilbur et al.(2009). *Journal of Archaeological*
626 *Science*, *36*(12), 2797-2804.
627

628 Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., & Rodrigo, A. G. (2003). Measurably
629 evolving populations. *Trends in ecology & evolution*, *18*(9), 481-488.
630

631 Drummond, A. J., Ho, S. Y., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating
632 with confidence. *PLoS biology*, *4*(5), e88.
633
634

635 Duchêne, S., Holmes, E. C., & Ho, S. Y. (2014). Analyses of evolutionary dynamics in viruses are
636 hindered by a time-dependent bias in rate estimates. *Proc. R. Soc. B*, *281*(1786), 20140732.
637

638 Duchêne, S., Geoghegan, J. L., Holmes, E. C., & Ho, S. Y. (2016 a). Estimating evolutionary rates
639 using time-structured data: a general comparison of phylogenetic methods. *Bioinformatics*, *32*(22),
640 3375-3379.

641 Duchêne, S., Holt, K. E., Weill, F. X., Le Hello, S., Hawkey, J., Edwards, D. J., ... & Holmes, E. C.
642 (2016 b). Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics*, *2*(11).
643

644 Duchene, S., Duchene, D. A., Geoghegan, J. L., Dyson, Z. A., Hawkey, J., & Holt, K. E. (2018).
645 Inferring demographic parameters in bacterial genomic data using Bayesian and hybrid phylogenetic
646 methods. *BMC evolutionary biology*, *18*(1), 95.
647

648

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., & Balloux, F. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature communications*, *6*, 7119.

Eldholm, V., Pettersson, J. H. O., Brynildsrud, O. B., Kitchen, A., Rasmussen, E. M., Lillebaek, T., ... & Alfsnes, K. (2016). Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, *113*(48), 13881-13886.

Folkvardsen, D. B., Norman, A., Andersen, Å. B., Michael Rasmussen, E., Jelsbak, L., & Lillebaek, T. (2017). Genomic Epidemiology of a Major Mycobacterium tuberculosis Outbreak: Retrospective Cohort Study in a Low-Incidence Setting Using Sparse Time-Series Sampling. *The Journal of infectious diseases*, *216*(3), 366-374.

Ford, C. B., Lin, P. L., Chase, M. R., Shah, R. R., Iartchouk, O., Galagan, J., ... & Flynn, J. L. (2011). Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature genetics*, *43*(5), 482.

Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., ... & Fortune, S. M. (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature genetics*, *45*(7), 784.

Glynn, J. R., Kremer, K., Borgdorff, M. W., Rodriguez, M. P., & Soolingen, D. V. (2006). Beijing/W genotype Mycobacterium tuberculosis and drug resistance.

Hanekom, M., Mata, D., van Pittius, N. G., van Helden, P. D., Warren, R. M., & Hernandez-Pando, R. (2010). Mycobacterium tuberculosis strains with the Beijing genotype demonstrate variability in virulence associated with transmission. *Tuberculosis*, *90*(5), 319-325.

Hershberg, R., Lipatov, M., Small, P. M., Sheffer, H., Niemann, S., Homolka, S., ... & Gagneux, S. (2008). High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. *PLoS biology*, *6*(12), e311.

Hershkovitz, I., Donoghue, H. D., Minnikin, D. E., Besra, G. S., Lee, O. Y., Gernaey, A. M., ... & Bar-Gal, G. K. (2008). Detection and molecular characterization of 9000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean. *PloS one*, *3*(10), e3426.

Ho, S. Y., Phillips, M. J., Cooper, A., & Drummond, A. J. (2005). Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular biology and evolution*, *22*(7), 1561-1568.

Ho, S. Y., Lanfear, R., Bromham, L., Phillips, M. J., Soubrier, J., Rodrigo, A. G., & Cooper, A. (2011). Time-dependent rates of molecular evolution. *Molecular ecology*, *20*(15), 3087-3101.

Holt, K. E., McAdam, P., Thai, P. V. K., Thuong, N. T. T., Ha, D. T. M., Lan, N. N., ... & Thwaites, G. (2018). Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nature genetics*, 1.

Kapur, V., Whittam, T. S., & Musser, J. M. (1994). Is Mycobacterium tuberculosis 15,000 years old?. *Journal of Infectious Diseases*, *170*(5), 1348-1349.

Kay, G. L., Sergeant, M. J., Zhou, Z., Chan, J. Z. M., Millard, A., Quick, J., ... & Achtman, M. (2015). Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nature communications*, *6*, 6717.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, *217*(5129), 624-626.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., ... & Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *science*, *288*(5472), 1789-1796.

Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing *Genome Research*, 22(3), 568-576.

Lee, R. S., Radomski, N., Proulx, J. F., Levade, I., Shapiro, B. J., McIntosh, F., ... & Behr, M. A. (2015). Population genomics of Mycobacterium tuberculosis in the Inuit. *Proceedings of the National Academy of Sciences*, *112*(44), 13609-13614.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60.

Li H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987-93.

Liu, Q., Ma, A., Wei, L., Pang, Y., Wu, B., Luo, T., ... & Gao, Q. (2018). China's tuberculosis epidemic stems from historical expansion of four strains of Mycobacterium tuberculosis. *Nature ecology & evolution*, 1.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20:1297-303.*

Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S. M., Rutaihwa, L. K., ... & Gagneux, S. (2018). Treemmer: a tool to reduce large phylogenetic data sets with minimal loss of diversity. *BMC bioinformatics*, *19*(1), 164.

Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., ... & Allix-Béguec, C. (2015). Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nature genetics*, *47*(3), 242.

Merker, M., Barbier, M., Cox, H., Rasigade, J. P., Feuerriegel, S., Kohl, T. A., ... & Andres, S. (2018). Compensatory evolution drives multidrug-resistant tuberculosis in Central Asia. *eLife*, 7:e38200.

Möller, S., du Plessis, L., & Stadler, T. (2018). Impact of the tree prior on estimating clock rates during epidemic outbreaks. *Proceedings of the National Academy of Sciences*, *115*(16), 4200-4205.

Morgan, G. J. (1998). Emile Zuckerkandl, Linus Pauling, and the molecular evolutionary clock, 1959–1965. *Journal of the History of Biology*, *31*(2), 155-178.

Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A., & Behr, M. A. (2002). Genomic deletions suggest a phylogeny for the Mycobacterium tuberculosis complex. *The Journal of infectious diseases*, *186*(1), 74-80.

Nicklisch, N., Maixner, F., Ganslmeier, R., Friederich, S., Dresely, V., Meller, H., ... & Alt, K. W. (2012). Rib lesions in skeletons from early neolithic sites in Central Germany: on the trail of tuberculosis at the onset of agriculture. *American journal of physical anthropology*, *149*(3), 391-404.

O'Neill, M. B., Shockey, A. C., Zarley, A., Aylward, W., Eldholm, V., Kitchen, A., & Pepperell, C. S. (2018). Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia. *bioRxiv*, 210161.

Paradis, E., Schliep, K., & Schwartz, R. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *1*, 3.

Pepperell, C. S., Casto, A. M., Kitchen, A., Granka, J. M., Cornejo, O. E., Holmes, E. C., ... & Feldman, M. W. (2013). The role of selection in shaping diversity of natural M. tuberculosis populations. *PLoS pathogens*, *9*(8), e1003543

Picard: https://github.com/broadinstitute/picard.

Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution*, *2*(1), vew007.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol*, *10*.

Ramsden, C., Melo, F. L., Figueiredo, L. M., Holmes, E. C., Zanotto, P. M., & VGDN Consortium. (2008). High rates of molecular evolution in hantaviruses. *Molecular Biology and Evolution*, *25*(7), 1488-1492.

Ribeiro, S. C., Gomes, L. L., Amaral, E. P., Andrade, M. R., Almeida, F. M., Rezende, A. L., ... & Lasunskaia, E. B. (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. *Journal of clinical microbiology*, JCM-00498.

Rieux, A., & Balloux, F. (2016). Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular ecology*, *25*(9), 1911-1924.

Rothschild, B. M., Martin, L. D., Lev, G., Bercovier, H., Bar-Gal, G. K., Greenblatt, C., ... & Brittain, D. (2001). Mycobacterium tuberculosis complex DNA from an extinct bison dated 17,000 years before the present. *Clinical Infectious Diseases*, *33*(3), 305-311.

Rutaihwa, L. K., Menardo, F., Stucki, D., Gygli, S. M., Ley, S. D., Malla, B., ... & Carter, J. E. (2018). Multiple Introductions of the Mycobacterium tuberculosis Lineage 2 Beijing into Africa over centuries. *bioRxiv*, 413039.

Seo TK, Thorne JL, Hasegawa M, Kishino H. A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times, Bioinformatics , 2002, vol. 18 (pg. 115-123)

SeqPrep: https://github.com/jstjohn/SeqPrep.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Steiner, A., Stucki, D., Coscolla, M., Borrell, S., & Gagneux, S. (2014). KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC genomics*, *15*(1), 881.

Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., ... & Gao, Q. (2016). Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nature genetics*, *48*(12), 1535.

Taylor, G. M., Murphy, E., Hopkins, R., Rutland, P., & Chistov, Y. (2007). First report of Mycobacterium bovis DNA in human remains from the Iron Age. *Microbiology*, *153*(4), 1243-1249.

Tong, K. J., Duchêne, D. A., Duchêne, S., Geoghegan, J. L., & Ho, S. Y. (2018). A comparison of methods for estimating substitution rates from ancient DNA sequence data. *BMC evolutionary biology*, *18*(1), 70.

To, T. H., Jung, M., Lycett, S., & Gascuel, O. (2015). Fast dating using least-squares criteria and algorithms. *Systematic biology*, *65*(1), 82-97.

Trewby, H., Wright, D., Breadon, E. L., Lycett, S. J., Mallon, T. R., McCormick, C., ... & Herzyk, P. (2016). Use of bacterial whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis. *Epidemics*, *14*, 26-35.

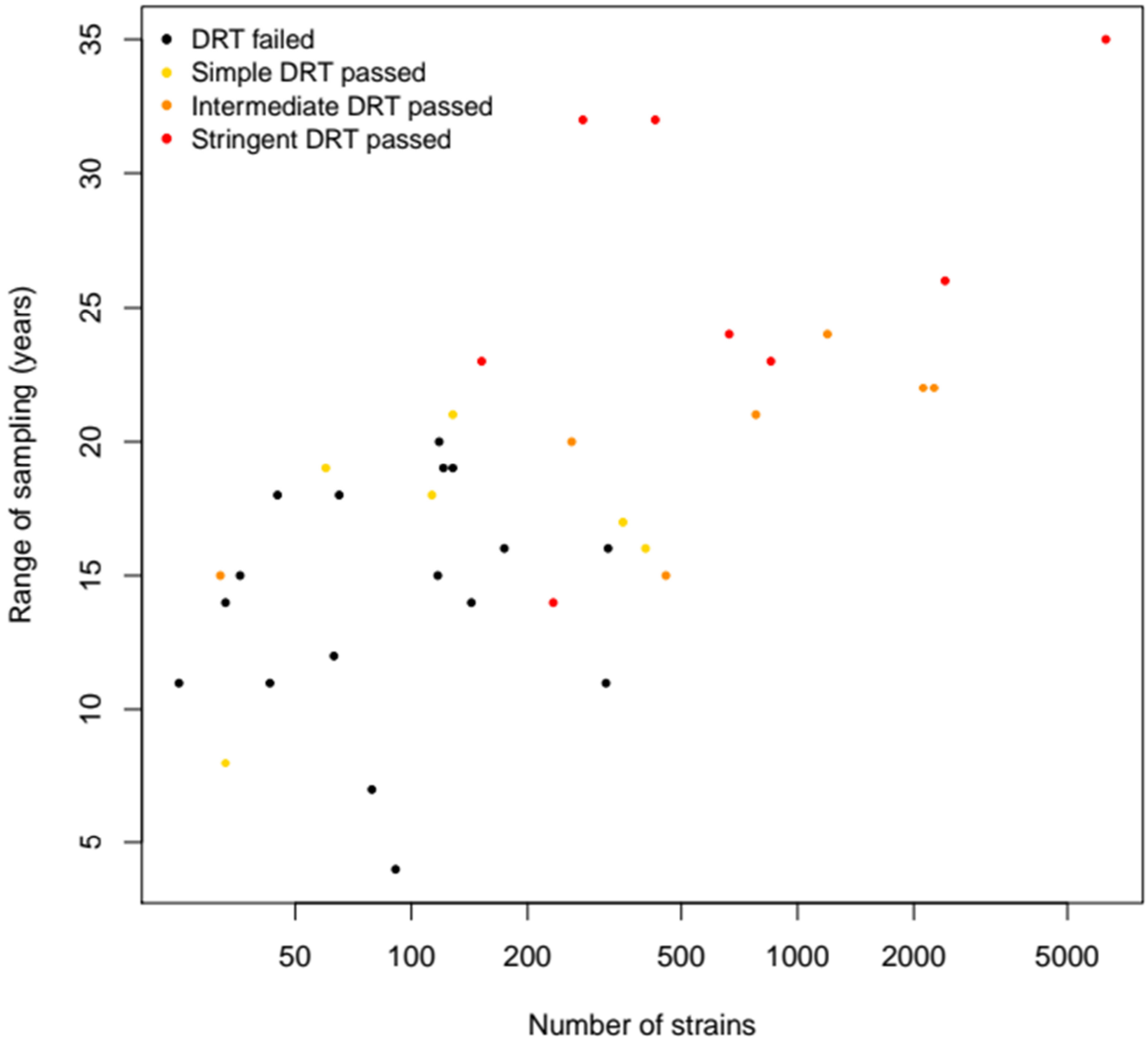
WHO. Global Tuberculosis Report (2018) http://www.who.int/tb/publications/global_report/en/

Wiens, K. E., Woyczynski, L. P., Ledesma, J. R., Ross, J. M., Zenteno-Cuevas, R., Goodridge, A., ... & Ray, S. E. (2018). Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. *BMC medicine*, *16*(1), 196.

Wilbur, A. K., Bouwman, A. S., Stone, A. C., Roberts, C. A., Pfister, L. A., Buikstra, J. E., & Brown, T. A. (2009). Deficiencies and challenges in the study of ancient tuberculosis DNA. *Journal of Archaeological Science*, *36*(9), 1990-1997.

Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., ... & Meyer, A. (2008). Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS pathogens*, *4*(9), e1000160.

Zuckerkandl, E., & Pauling, L. (1962). Molecular disease, evolution and genetic heterogeneity.



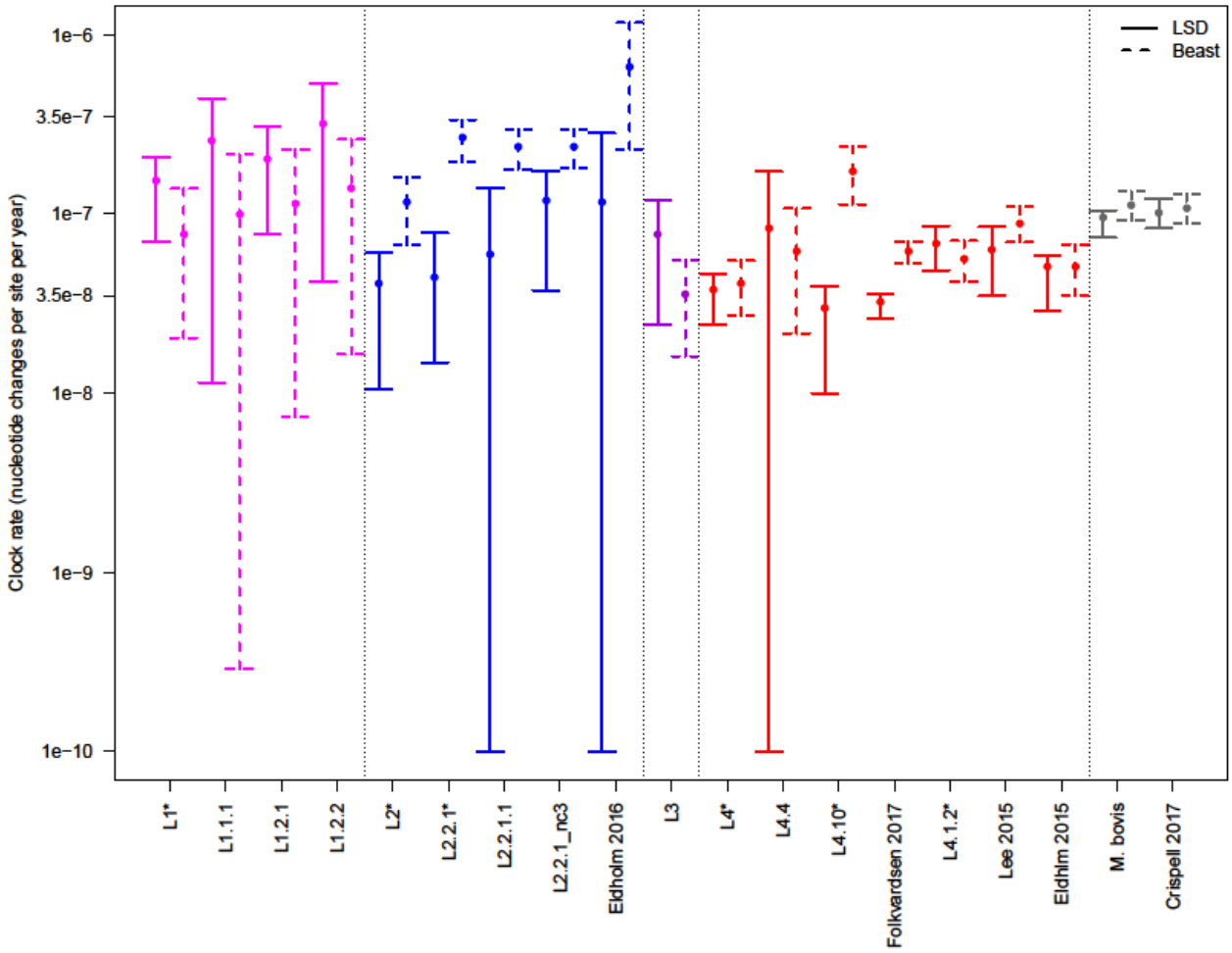
847

848

849 **Figure 1.** Results of the DRT for all data sets ordered by size and temporal range. Data sets with fewer

850 strains sampled in a shorter period of time tended to fail the DRT.

851



852

853 **Figure 2.** Estimated clock rates of all lineages, sub-lineages and local data sets that passed the DRT.

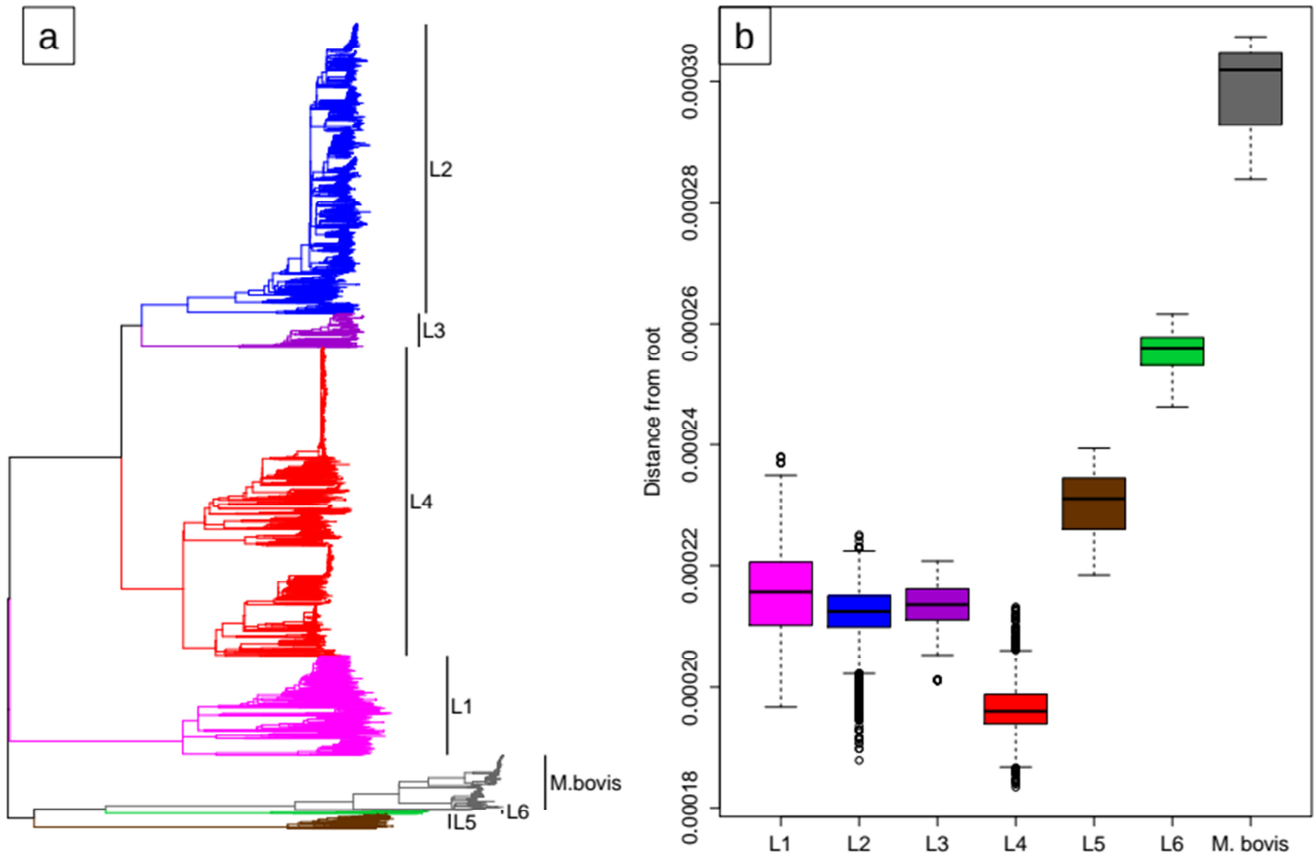
854 Solid lines represent the 95% confidence interval estimated with LSD, dashed lines represent the 95%

855 highest posterior density (HPD) estimated by Beast (the larger dot is the median of the posterior

856 distribution). Data sets marked with * have been reduced to 500 randomly picked strains for the Beast

857 analysis.

858



859

860

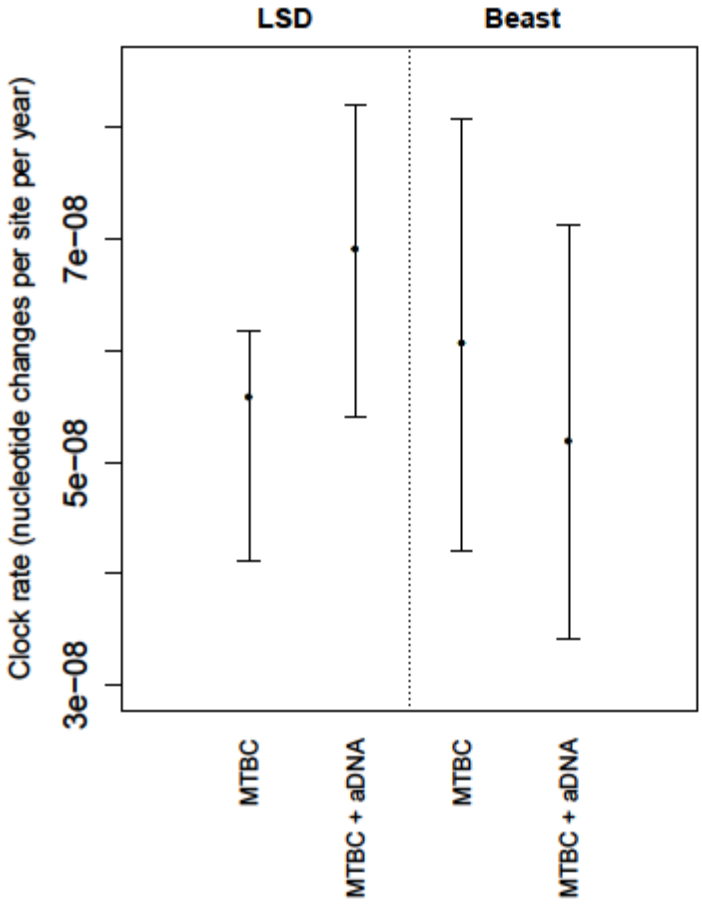
861 **Figure 3. a)** The Maximum Likelihood tree of the 6,285 strains considered in this study rooted with the

862 genome sequence of *M. canetti*. **b)** Phylogenetic distance from the root (expected nucleotide changes

863 per site) of MTB strains by lineage

864

865



866

867

868 **Figure 4.** Comparison of the MTB complex clock rates with and without the 3 aDNA samples, the

869 LSD analysis included 6,285 strains, and the tree was rooted with *M. canettii*, in the Beast analysis we

870 included 500 randomly selected strains and no outgroup.

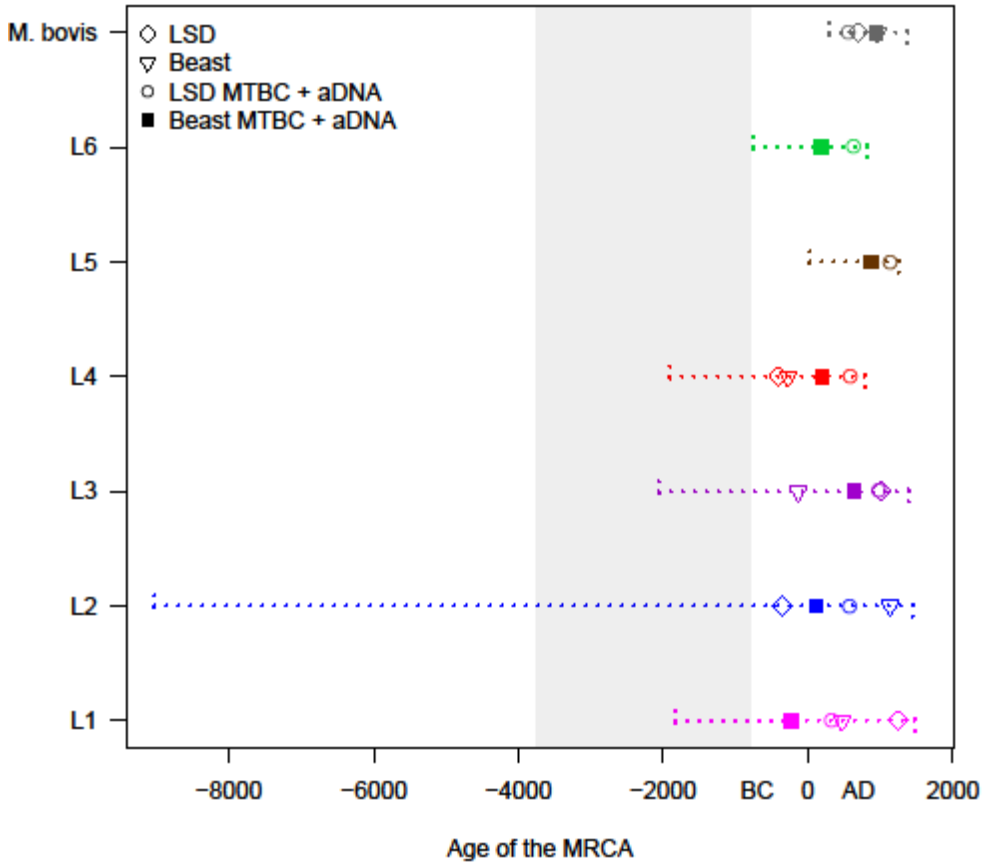
871

872

873

874

875



876
877
878
879
880
881
882
883
884

885 **Figure 5.** The inferred age of the MTB lineages.

886 LSD : results of the LSD analysis performed on the individual lineages. Beast: results of the Beast
887 analysis performed on the individual lineages (median values). LSD MTBC + aDNA: results of the
888 LSD analysis performed on the complete data set of 6.285 strains + 3 aDNA samples, the age of the
889 MRCA of the individual lineages was identified on the calibrated tree. Beast MTBC + aDNA: results of
890 the Beast analysis performed on the random subsample of 500 strains + 3 aDNA samples, the age of the
891 MRCA of the individual lineages was identified on the calibrated tree (median values). The confidence
892 intervals were obtained merging the 95% CI and HPD of all analyses. The shaded area represents the
893 age of the MRCA of the MTB complex obtained with the LSD analyses (with and without aDNA, the
894 two 95% CI were merged). For L5 and L6 we report only the age inferred on the complete MTB
895 complex tree, because when analyzed individually these two data sets showed a lack of temporal
896 structure (they failed the DRT).

897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913

SI Appendix

The molecular clock of *Mycobacterium tuberculosis*

Comparison of sub-sampling strategies

Several data sets considered in this study are too large for Bayesian phylogenetic analysis with Beast. Therefore we decided to sub-sample the largest data sets to a size of 500 strains (MTBC, L1, L2, L2.2.1, L4, L4.1.2, L4.10). The only exception was Folkvardsen 2017, this data set is composed of 853 strains, but it has less than 1,000 polymorphic positions and it was computationally treatable by Beast. We adopted two different sub-sampling strategies: 1) random selection and 2) we used Treemmer, an algorithm that reduces the size of phylogenetic data set maintaining a representative diversity (Menardo et al. 2018). In this section we compare the results of these two sampling strategies.

First we repeated the DRT on the reduced datasets: we found that four of the seven randomly selected sub-sets and three of the seven trimmed data sets passed the DRT (Fig. S8). We compared the LSD estimates of the complete and reduced data sets: four of the random sub-sets produced higher estimates than the complete data sets, and all seven trimmed data sets resulted in higher estimates compared to the complete ones, although only marginally, and with overlapping confidence intervals (Fig. S9). The only exception was the MTB complex, for which the rate estimates of the random and trimmed data sets were not overlapping among them, additionally they were both not overlapping with the estimates obtained with complete data set. This indicates that the strong subsampling of the large MTB complex, from 6,285 to 500 strains, results in moderately different results regardless of the sampling strategy, and that this effect is stronger in the trimmed sub-sample. The reason for the apparent upward bias of the clock rate of trimmed data sets, could be that Treemmer tends to prune tips with shorter terminal branches, and this could influence the molecular clock analysis, especially if the original tree is pruned

940 heavily. If we exclude the MTBC data set this effect appear to be modest in magnitude, and in one case,
941 L2, the trimmed data sets produced point estimates lower than the random sub-samples (Figs. S8-S9).
942 We then performed the Bayesian analysis with Beast on all reduced data set and compared the results
943 between random sub-samples and trimmed sub-sampled (Fig. S9). Importantly all reduced data sets
944 when analyzed with Beast had a clock rate posterior distribution distinct from the prior, indicating the
945 presence of temporal structure, even though some of them did not pass the DRT (Figs. S2-S5, Table
946 S2). The trimmed sub-samples produced higher clock rates in five data sets, while the random sub-
947 samples resulted in higher clock rates in two data sets. Importantly, for three data sets the beast
948 estimates of random and trimmed sub-samples had non-overlapping 95% HPD: L1 and L4 had larger
949 clock rates for the trimmed sub-samples, and L4.10 had a larger clock rate for the random sub-sample
950 (Fig. S9, Table S2), we will discuss below what could cause those large differences.
951 Overall we found that the sub-sampling strategy can influence the results of molecular clock analyses,
952 and therefore should be considered carefully. This was especially true with strong sub-sampling of
953 large data sets, or when using Beast. It appears that there is not an optimal strategy valid for all data
954 sets, when the original sampling is highly biased, with many strains sampled from the same outbreak or
955 from a few outbreaks (i.e L4.1.2 and L4.10), and therefore highly similar, it is advisable to use methods
956 like Treemmer to balance the data set. However, if a very large tree is reduced with Treemmer the
957 resulting data set could be enriched with long terminal branches, and lead to an overestimation of the
958 clock rate. In some cases more complex strategies that preserve phylogenetic diversity and a broad
959 sampling period might be necessary. Additionally in a Bayesian analysis it is not obvious how to
960 choose the tree prior of data sets reduced by Treemmer: the standard coalescent-based tree priors
961 implemented in Beast assume random sampling from a single population, and this condition is not
962 respected by Treemmer, however in most cases also the complete data set is not a random sample of a
963 population.

964 **The data sets L4_nc and Eldholm 2016**

965 The data set L4_nc, resulted in an estimated rate of 6.89×10^{-7} and 3.54×10^{-6} with LSD and Beast
966 respectively (Fig. S13, Methods); and the data set from Eldholm 2016, resulted in an estimated rate of
967 1.17×10^{-7} and 6.63×10^{-7} with LSD and Beast respectively (Fig. S14). These estimates, especially for
968 L4_nc, are much higher than the results of all other data sets included in this study, and compare with
969 fast evolving bacteria like *Enterococcus faecium* and *Staphylococcus aureus* (Duchene et al. 2016 b).
970 However there are several reasons to doubt their accuracy: 1) L4_nc and Eldholm 2016 are the smallest
971 among all considered data sets, with 32 and 33 strains respectively. 2) Most strains in both data sets are
972 identical or nearly so, and collected in the same year or in a few years (Fig. S6, S13 and S14). 3) The
973 root to tip regressions and the very high coefficient of variations (COVs) compared to all other data
974 sets, suggest that the temporal signal comes from one single strain in L4_nc, and from just a few strains
975 in Eldholm 2016 (Figs. S3 and S4, Table S2). 4) In their original analysis of the Afghan family (the
976 Eldholm 2016 data set), Eldholm and colleagues (Eldholm et al. 2016) performed several Bayesian
977 analyses and estimated a clock rate of $\sim 3 \times 10^{-7}$, with a slightly larger sample size (35 strains).

978

979 **Comparison LSD – Beast**

980 It was shown before that although LSD and Beast produce comparable results, analyses with a relaxed
981 clock tend to produce higher clock rates than strict clocks (Duchene et al. 2016). To et al. (2015) found
982 that in some cases Beast overestimated the clock rates compared to LSD, but this bias was not visible
983 on the age of the MRCA. Additionally, Möller et al. (2018) showed that standard tree priors used in
984 Beast can lead to the overestimation of clock rates in presence of population structure. Since we also
985 observed some differences between the results of LSD and Beast, we did a systematic comparison to
986 identify eventual biases. For this analysis we considered all data sets that passed the simple DRT (20

987 data sets). We found no systematic bias among the two methods (Table S5), Beast resulted in higher
988 clock rates for 10 of the 20 data sets, and in older MRCA for 11 data sets. The median proportional
989 distance between methods was 3.5% and 5.7% for clock rates and age of MRCAs respectively, this is
990 less than the distance between LSD and Beast found by Duchene et al. (2016) on viruses data sets.
991 Nevertheless the total overlap between the 95% CI of LSD and the 95% HPD of Beast was only around
992 40%, and about a quarter of all data sets had no overlap. Interestingly, two of the five data sets that had
993 non-overlapping clock rate estimates, L4.10 and Folkvardsen 2017, had a very high COVs, 2.9 and 1.2
994 respectively, indicating an important among lineage variation of the clock rate (Table S2). This
995 supports the finding of Duchene et al. (2016), that a high rate variation among lineages is correlated
996 with larger differences between the results of LSD and Beast.
997 One important difference between LSD and Beast is that the first is not sensitive to the tree shape and
998 to the degree of phylogenetic and temporal clustering, while Beast assumes a tree prior (coalescent
999 constant population size in this case). This observation is important to understand the different Beast
1000 results for the two sub-sampling strategies. For example the data set L4.10 has 1,195 strains, 853 of
1001 them belong to a single outbreak in Denmark (the Folkvardsen 2017 data set), and are very similar to
1002 each other. Sub-sampling randomly, preserves the original shape of the tree, most strains will belong to
1003 the Folkvardsen 2017 data set. Conversely, Treemmer will prune the Danish outbreak heavily, because
1004 it is composed of very similar strains, and the resulting tree will have a drastically different shape. The
1005 finding that the Beast results for these two sub-samples are very different, while the LSD results are
1006 comparable confirms that the tree shape can influence the results of Beast. We mentioned before that
1007 prior misspecification can lead to biased results (Möller et al. 2018), while for local populations such as
1008 outbreaks it is possible that one of the tree priors available in Beast captures the processes generating
1009 the tree topology, this is very unlikely for more diverse data sets. In many cases the tree topology is the

1010 results of complex population dynamics including demographic fluctuations, population subdivisions
1011 and biased sampling.
1012

1013 **Sensitivity to the tree prior**

1014 Choosing the tree prior is a critical aspect of Bayesian phylogenetics, here we investigate the effect of
1015 different tree priors on the results of Beast by repeating it assuming a coalescent tree prior with
1016 exponential population growth. For this analysis we considered the five outbreaks that passed the DRT
1017 (Eldholm et al. 2015, Lee et al. 2015, Eldholm et al. 2016, Crispell et al. 2017, Folkvardsen et al.
1018 2017), these are the only data sets that possibly represent a random sampling of a single population (an
1019 assumption of all coalescent-based tree prior). The constant population prior is a specific case of the
1020 exponential growth prior (when the growth rate = 0), therefore if the 95% HPD of the growth rate does
1021 not include 0, we can conclude that the data support a demographic model with variable population
1022 size.

1023 We found that the HPD of the growth rate of two data sets, Lee 2015 and Eldholm 2016, includes 0 and
1024 support a stable population size, for these two data sets the estimates of clock rates and the age of the
1025 MRCA are essentially identical when using the different priors (Table S3). The other three data sets,
1026 Eldholm 2015, Crispell 2017 and Folkvardsen 2017, had a growth rate greater than 0, thus supporting a
1027 growing population size. These findings confirm the results of the original publications, where
1028 Bayesian model selection supported models with population growth (Eldholm e al. 2015, Crispell et al.
1029 2017, Folkvardsen et al. 2017). Interestingly for these data sets the clock rates estimates obtained with
1030 the exponential growth prior were always larger compared to the results obtained with the constant
1031 population size prior (Table S3). More importantly we obtained a much younger age of the MRCA with
1032 the exponential population growth prior compared to the constant population size. The magnitude of

1033 the differences among the results obtained with different priors correlates with the growth rate (Table
1034 S3). For the slowly growing New Zealand population of *M. bovis* (Crispell 2017) we obtained largely
1035 overlapping HPDs of the age of the MRCA. For the Danish outbreak (Folkvardsen 2017), which had a
1036 higher growth rate, the HPD of the age of the MRCA were only barely overlapping. And finally, for the
1037 Argentinian outbreak (Eldholm 2015), the data sets with the larger estimated growth rate, the age of the
1038 MRCA obtained with the exponential growth prior was three times younger compared to the results of
1039 the constant population size prior, and the HPDs were not overlapping.

1040 Overall these results show that the tree prior can have a large influence on the results of molecular
1041 clock analyses, especially when the goal is to the date the tree. Therefore it is crucial to try different
1042 tree priors even if we are not interested in the demographic history. When they produce diverging
1043 results one should perform Bayesian model selection. This is more feasible for local populations and
1044 outbreaks, for more complex data sets it is often impossible to avoid tree prior misspecification, and we
1045 address this point in the Discussion.

1046

1047

1048

1049

1050

1051

1052

1053

1054

Supplementary Tables and Figures

Supplementary Table S1

File: Supplementary_table1.txt

List of strains used in this study with sampling year and accession numbers

Supplementary Table S2

File: Supplementary_table2.xlsx

Results of Beast and LSD for all data sets

1079 **Supplementary Table S3**

1080 Comparison of Beast results with different tree priors

Data set	Tree prior[1]	MRCA[2]	Clock rate[3]	COV[4]	Growth rate[5]
Lee 2015	Exponential	1919.4 [1887.4, 1943.0]	8.87E-08 [6.97E-8, 1.09E-7]	0.30 [2.19E-4, 0.59]	-0.01 [-0.02, 0.01]
Lee 2015	Constant	1920.3 [1890.1, 1945.1]	8.89E-08 [7.00E-8, 1.10E-7]	0.31 [1.62E-4, 0.59]	-
Eldholm 2015	Exponential	1976.3 [1963.4, 1985.6]	6.57E-08 [4.62E-8, 8.55E-8]	0.83 [0.55, 1.145]	0.24 [0.14, 0.37]
Eldholm 2015	Constant	1892.2 [1783.7, 1952.6]	5.08E-08 [3.50E-8, 6.76E-8]	0.97 [0.64, 1.35]	-
Eldholm 2016	Exponential	1989.4 [1949.9, 2005.9]	6.73E-07 [2.43E-7, 1.20E-6]	2.20 [0.82, 4.34]	-0.04 [-0.19, 0.1]
Eldholm 2016	Constant	1989.4 [1951.3, 2005.3]	6.63E-07 [2.3E-7, 1.17E-6]	2.37 [1.02, 4.60]	-
Folkvardsen 2017	Exponential	1971.4 [1961.4, 1979.6]	7.20E-08 [6.45E-8, 8.00E-8]	0.98 [0.77, 1.22]	0.11 [0.09, 0.13]
Folkvardsen 2017	Constant	1934.1 [1871.1, 1967.5]	6.16E-08 [5.33E-8, 7.01E-8]	1.16 [0.86, 1.49]	-
Crispell 2017	Exponential	1581.6 [1428.2, 1713.1]	1.16E-07 [9.46E-8, 1.38E-7]	0.49 [0.40, 0.59]	0.01 [3.14E-3, 0.01]
Crispell 2017	Constant	1427.8 [1185.2, 1617.0]	1.08E-07 [8.86E-8, 1.29E-7]	0.50 [0.40, 0.60]	-

1081
1082
1083 [1] Coalescent tree prior with constant population size or exponential population growth / shrinkage
1084 [2] Age of the most recent common ancestor (years AD), median value and 95% HPD
1085 [3] Clock rate in expected nucleotide changes per year, median value and 95% HPD
1086 [4] Coefficient of variation, median value and 95% HPD
1087 [4] Exponential growth rate, median value and 95% HPD
1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098 **Supplementary Table S4**

1099 The age of the MTB complex and of its lineages resulted from different analyses

1100

Lineage	LSD[1]	Beast[2]	LSD + aDNA[3]	Beast + aDNA[4]
MTBC	-2287 [-3742, -1849]	NA	-1449 [-2414, -812]	NA
L1	1245 [510, 1470]	468 [-1833, 1302]	327	-230 [-1354, 489]
L2	-358 [-9035, 587]	1134 [498, 1442]	575	114 [-809, 732]
L3	1006 [-949, 1385]	-136 [-2068, 853]	983	638 [-55, 1111]
L4	-410 [-1906, 91]	-277 [-1225, 426]	582	200 [-676, 779]
L5	NA	NA	1131	877 [29, 1250]
L6	NA	NA	633	181 [-766, 809]
M. bovis	692 [297, 849]	964 [382, 1377]	551	936 [374, 1286]

1101
1102
1103
1104
1105

1106 [1] Age of the most recent common ancestor (negative values = BC, positive values = AD), point estimate
1107 and 95% CI. These estimates refer to the individual analyses performed on each data set.
1108 [2] Age of the most recent common ancestor (negative values = BC, positive values = AD), median value
1109 and 95% HPD. These estimates refer to the individual analyses performed on each data set. Since Beast
1110 placed the root in the "wrong" position we have no estimates for the MTB complex.
1111 [3] Age of the most recent common ancestor (negative values = BC, positive values = AD), point estimate
1112 and 95% CI. These estimates refer to a single analysis performed on the complete data set of 6,285
1113 strains + 3 aDNA samples; LSD outputs confidence intervals only for the MRCA of the tree and not for
1114 the nodes.
1115 [4] Age of the most recent common ancestor (negative values = BC, positive values = AD), median value
1116 and 95% HPD. These estimates refer to a single analysis performed on the random subset of MTBC
1117 composed of 500 strains + 3 aDNA samples; since Beast placed the root in the "wrong" position we
1118 have no estimates for the MTB complex. Nevertheless we could retrieve the age of the MRCA of the
1119 individual lineages.
1120

1121

1122

1123

Supplementary Table S5

Comparison of Beast and LSD results

	Bias[1]	**Distance**[2]	**Overlap**[3]	**Not overlapping**[5]
Clock rate	55%	3.5%	39.2%	5
Age of MRCA	45%	5.7%	41.6%	2

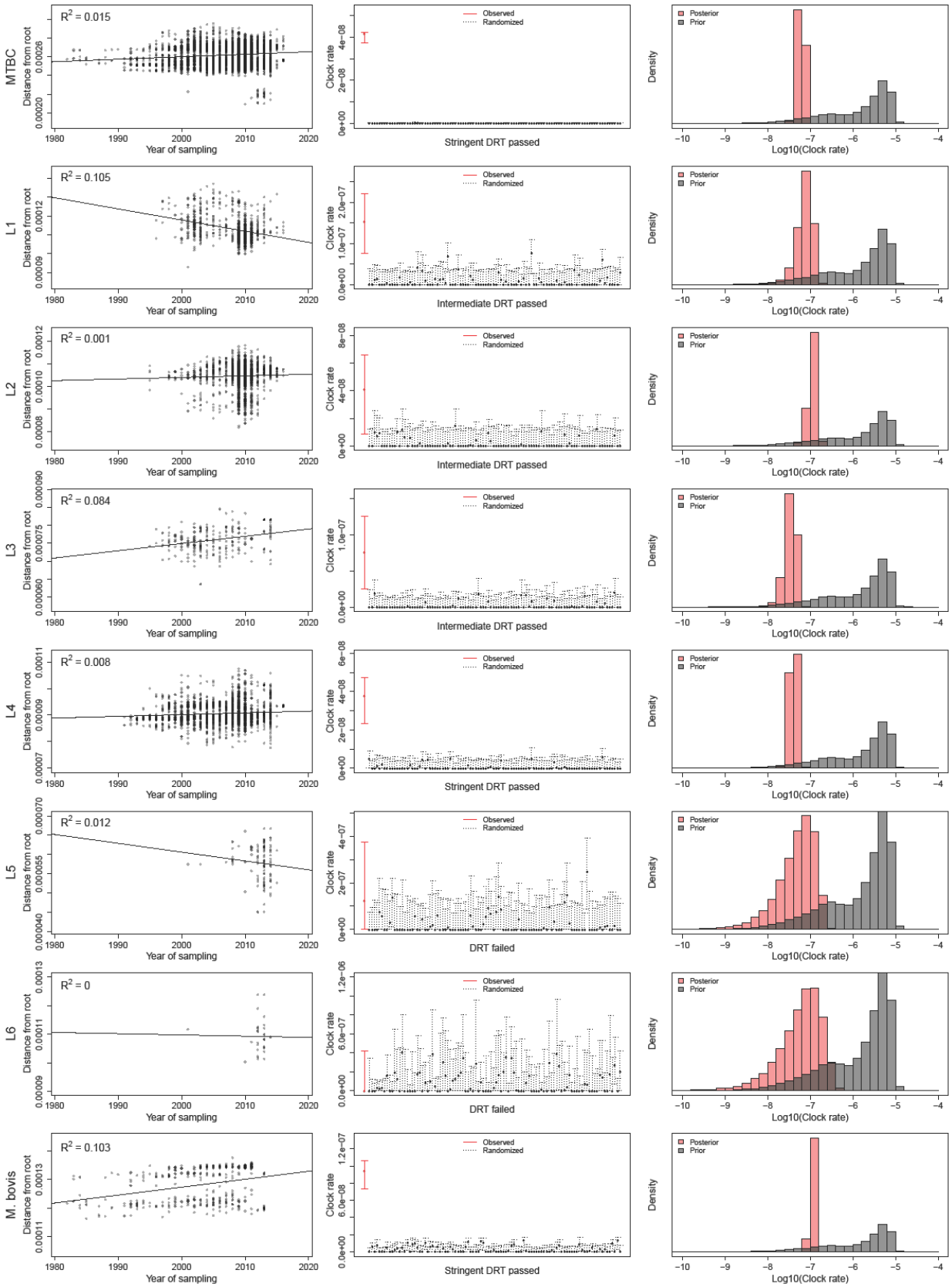
[1] Percentage of data sets with a higher Beast estimate
[2] Median proportional distance between estimates, calculated as:
(Beast estimate – LSD estimate) / Beast estimate * 100
[3] Percentage of overlapping CI / HPD
[4] Number of data sets with not overlapping CI /HPD

Supplementary Table S6

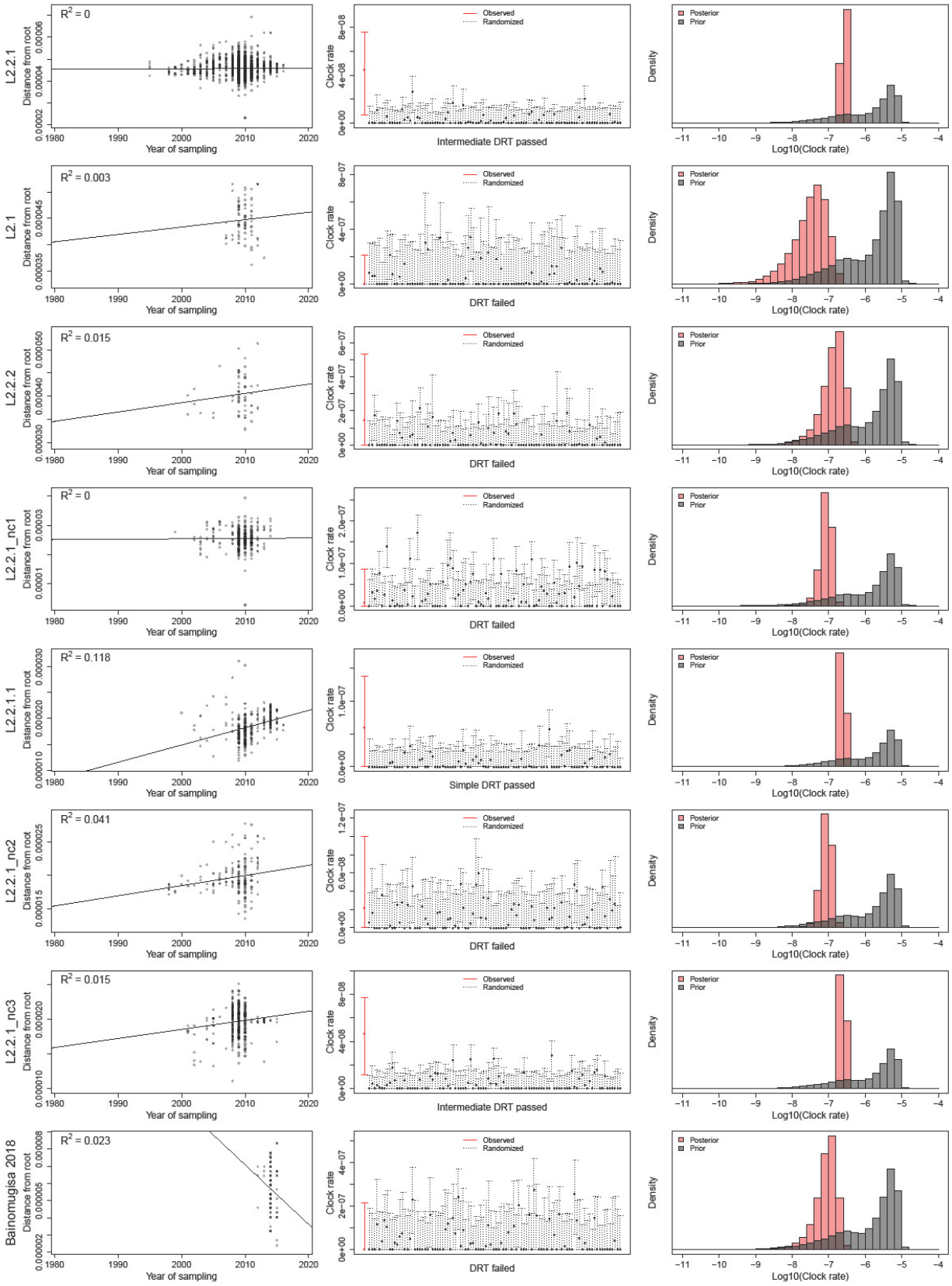
File: Supplementary_table6.txt

List of all accession numbers, before filtering

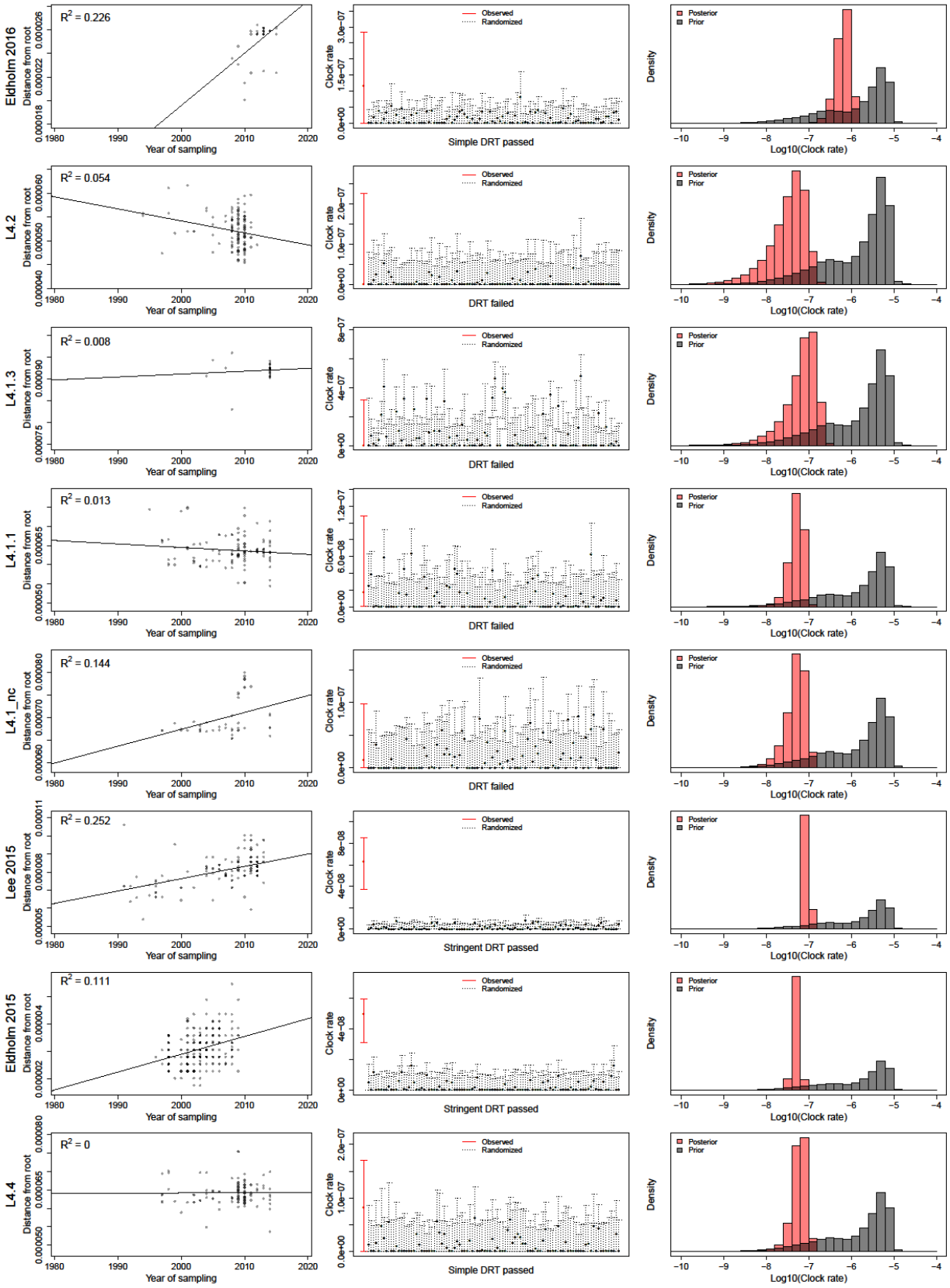
Supplementary Figure S1



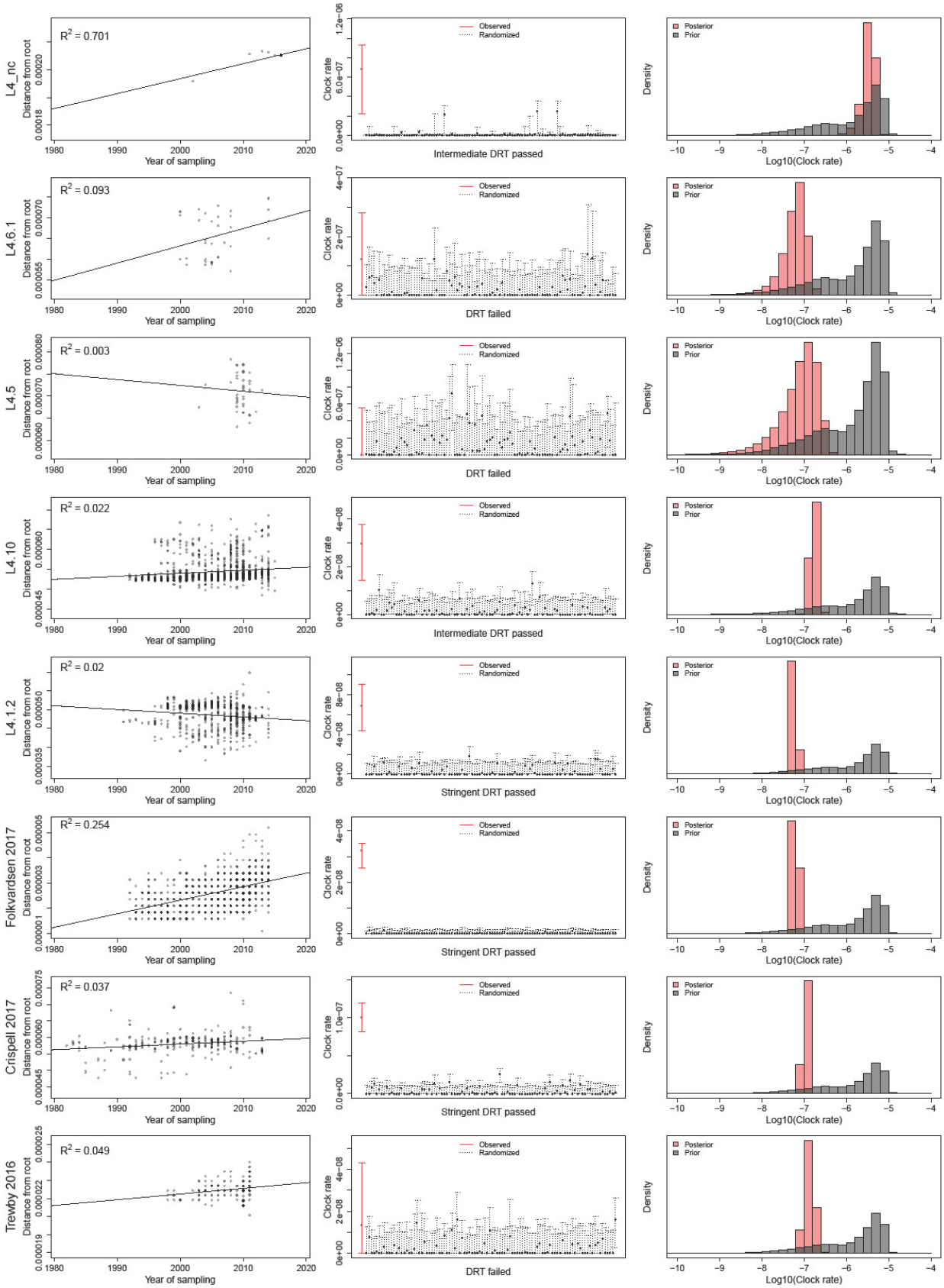
Supplementary Figure S2



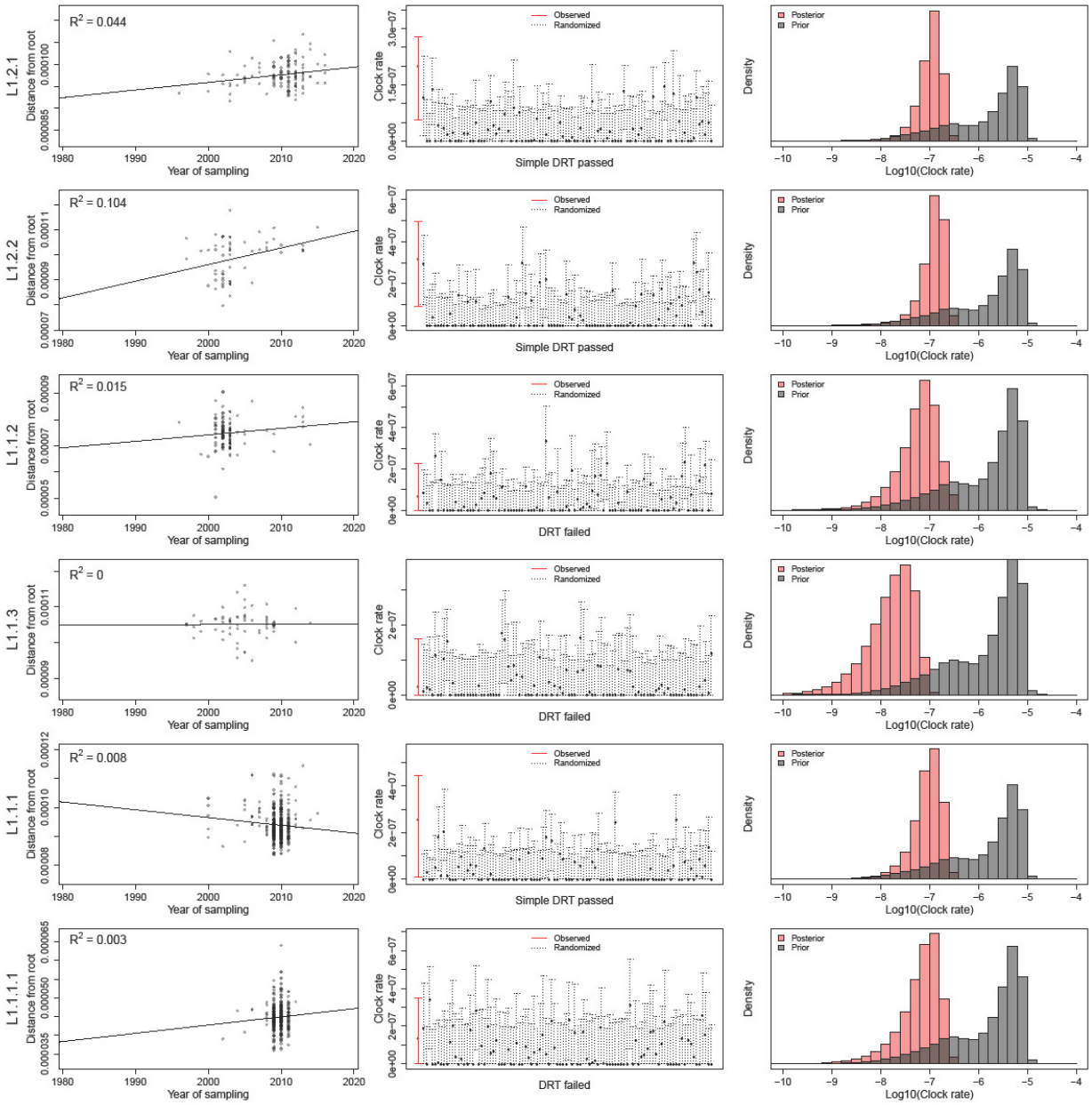
Supplementary Figure S3



Supplementary Figure S4



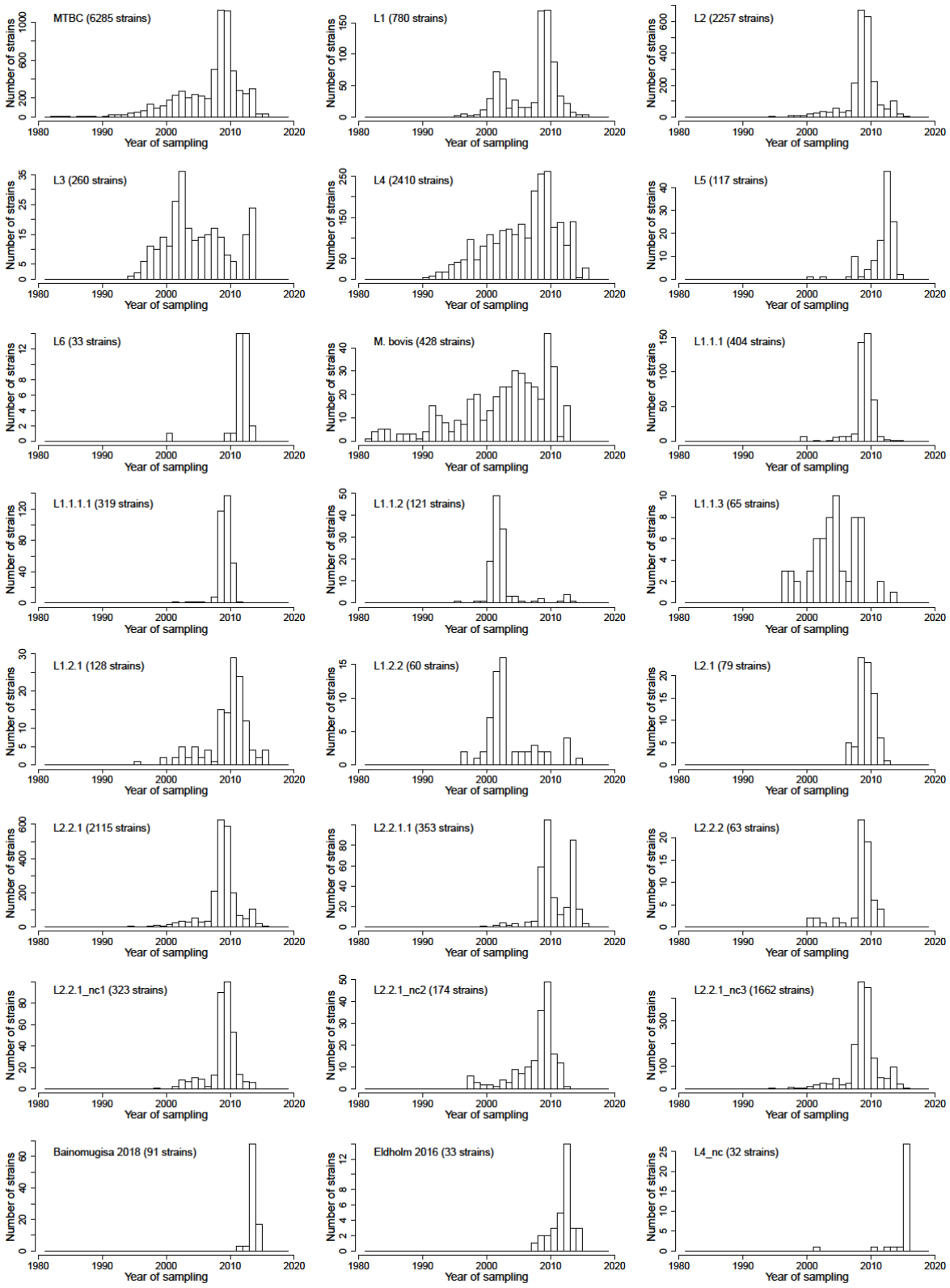
Supplementary Figure S5



Supplementary Figures S1-S5

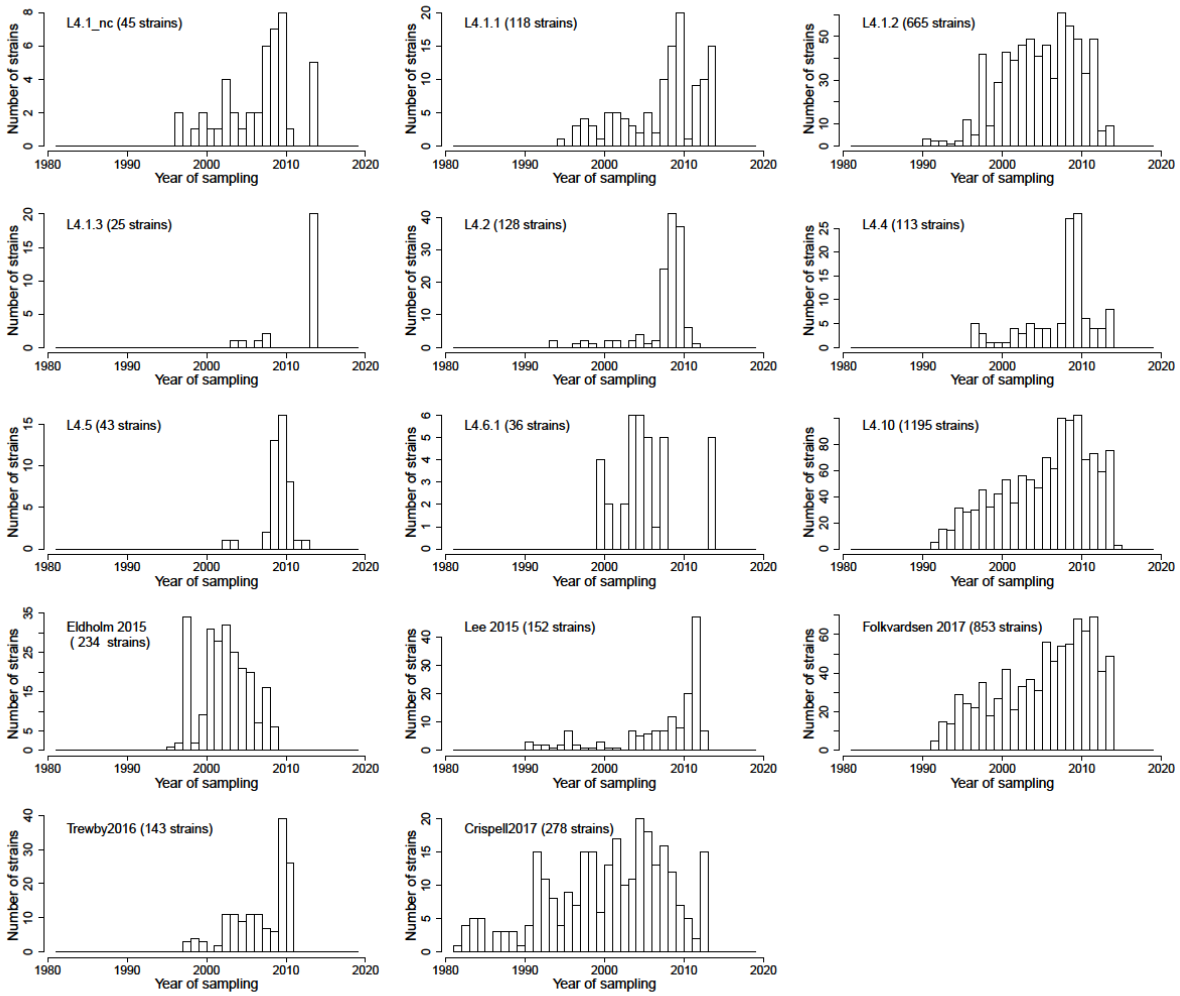
For each data set we report the results of the root to tip regression, the results of the Date Randomization Test (DRT) with LSD, and the comparison of the prior and posterior distribution of the clock rate. The simple DRT is passed when the clock rate estimate for the observed data does not overlap with the range of estimates obtained from the randomized sets. The intermediate DRT is passed when the clock rate estimate for the observed data does not overlap with the confidence intervals of the estimates obtained from the randomized sets. The stringent DRT is passed when the confidence interval of the clock rate estimate for the observed data does not overlap with the confidence intervals of the estimates obtained from the randomized sets. Large data sets (MTBC, L1, L2, L2.2.1, L4, L4.10, L4.1.2) were randomly sub-samples to 500 strains for the Beast analysis.

1191 **Supplementary Figure S6**



1192

1193 **Supplementary Figure S7**



1194

1195 **Supplementary Figures S6-S7**

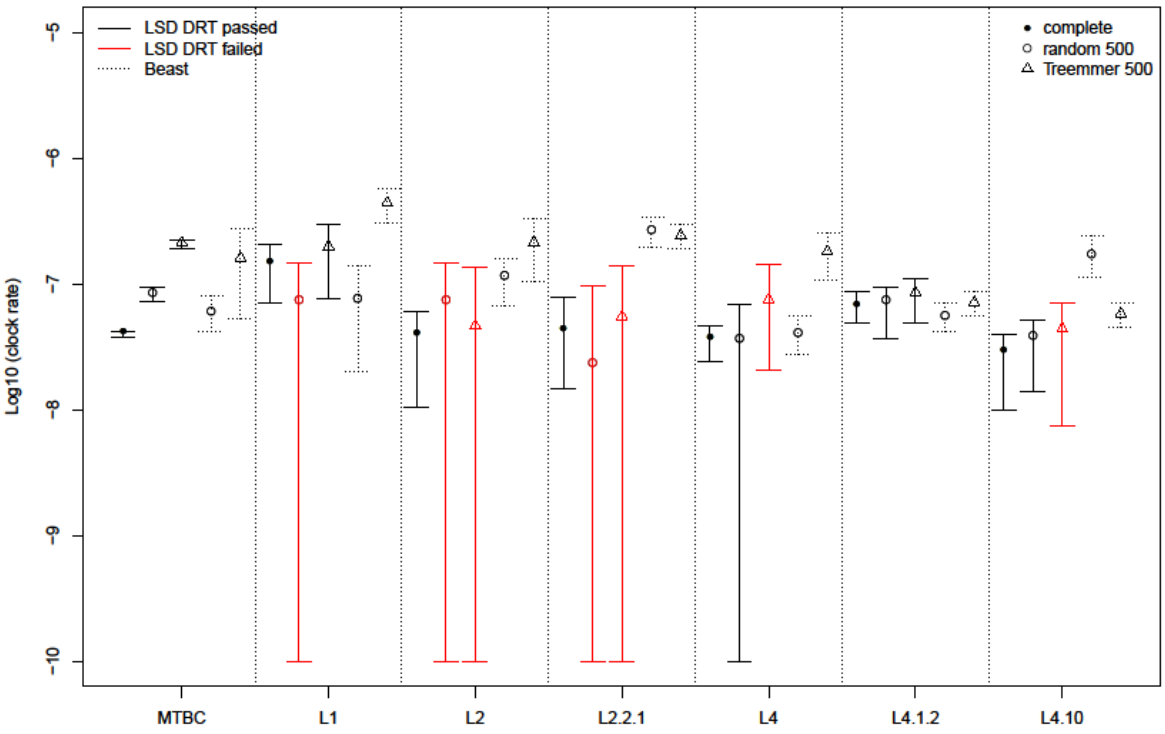
1196 Distribution of the sampling years for all data sets

1197

1198

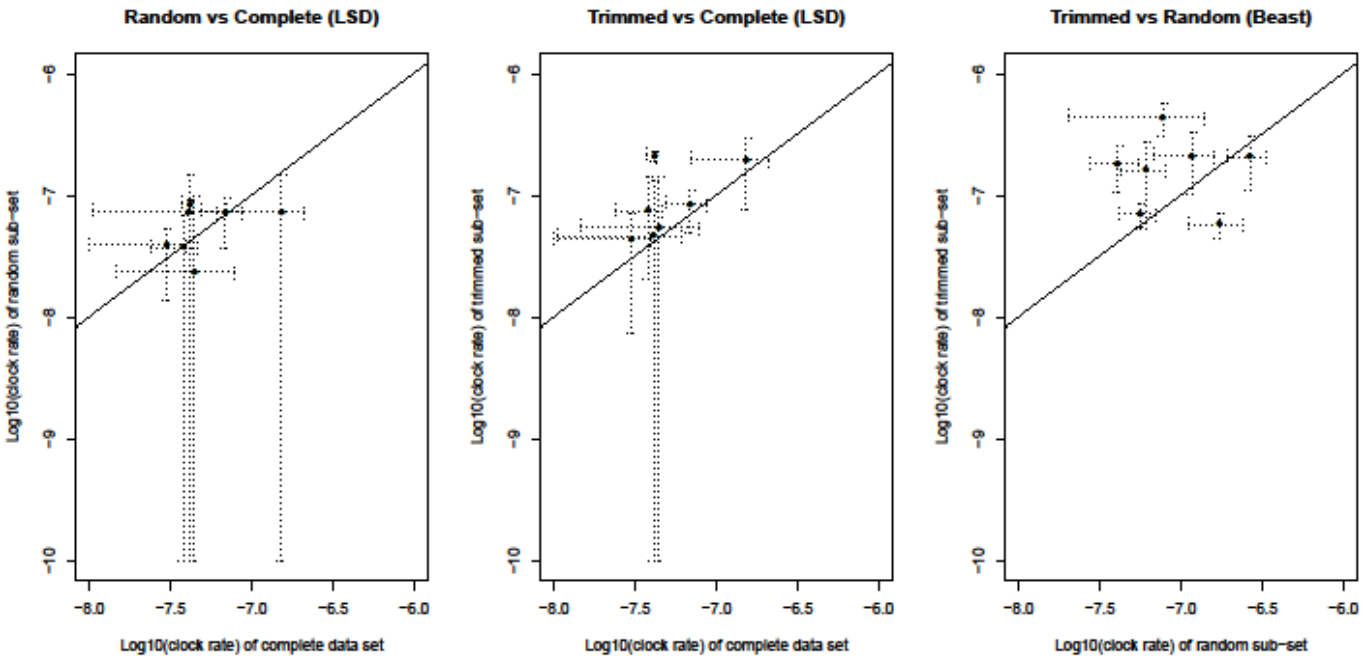
1199

1200



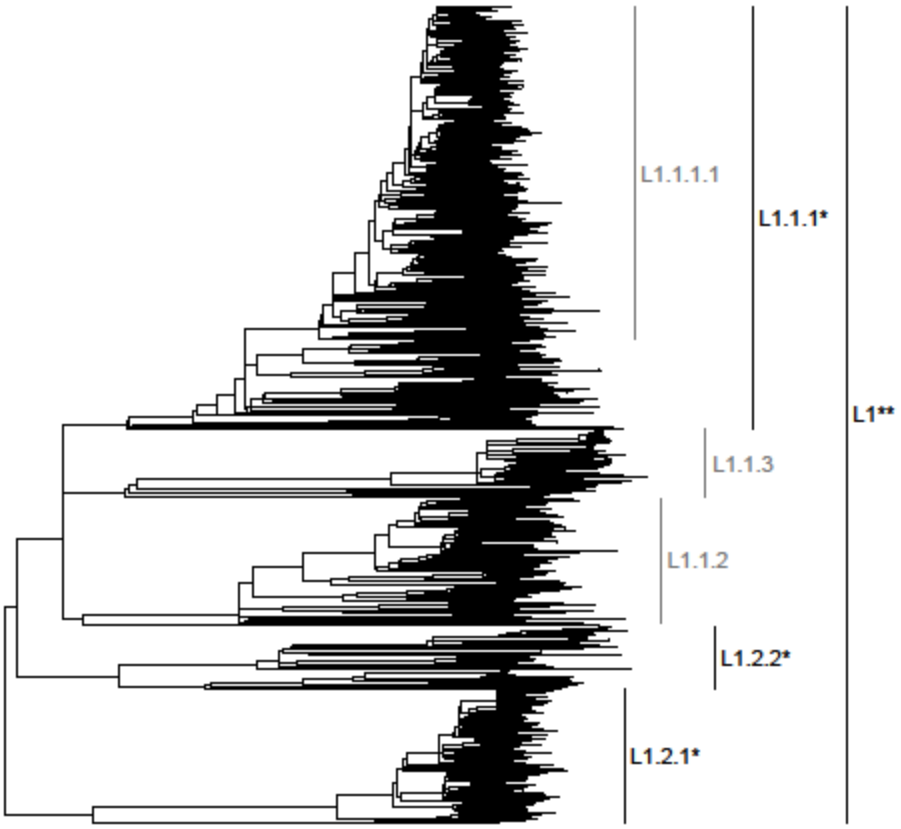
Supplementary Figure S8

Comparison of results for the seven largest data sets. For each data set we reported the LSD results on the complete data set, and the Beast and LSD results on the two sub-sampled data sets (random sub-sample and sub-sample selected with Treemmer)



Supplementary Figure S9

Comparison of results of sub-samples with complete data sets with LSD, and among sub-samples with Beast



1224

1225 **Supplementary Figure S10**

1226 Sub-lineages of L1 that were included in the analysis. Clades in colored in gray did not pass the DRT,
1227 clades colored in black passed the DRT. *: simple DRT passed, ** intermediate DRT passed, ***:
1228 stringent DRT passed.

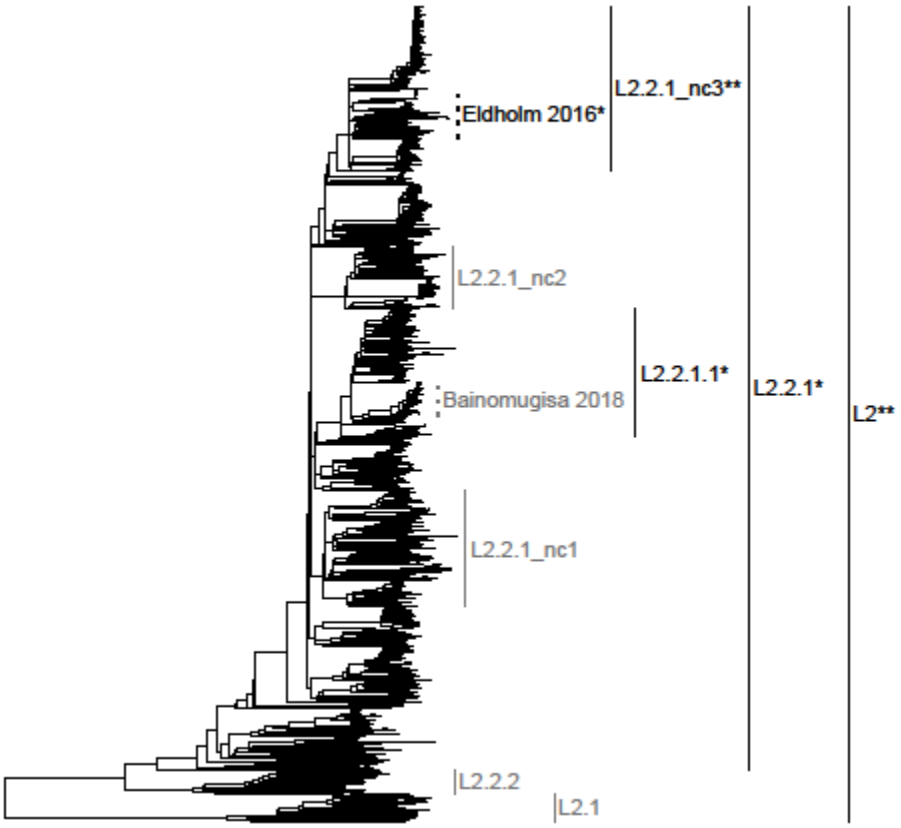
1229

1230

1231

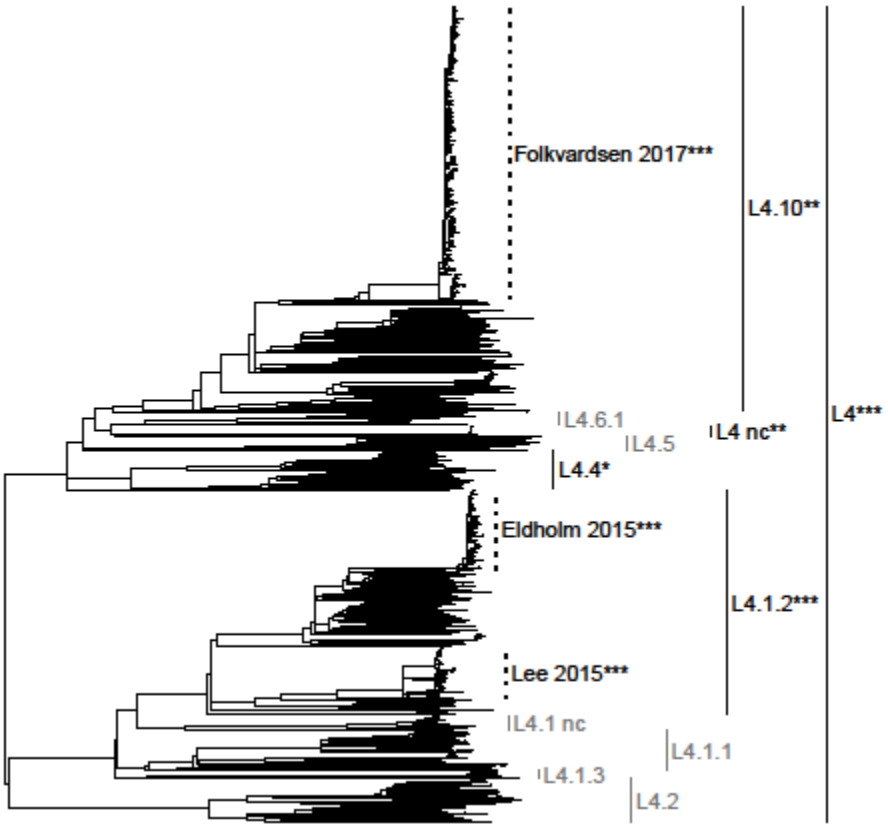
1232

1233



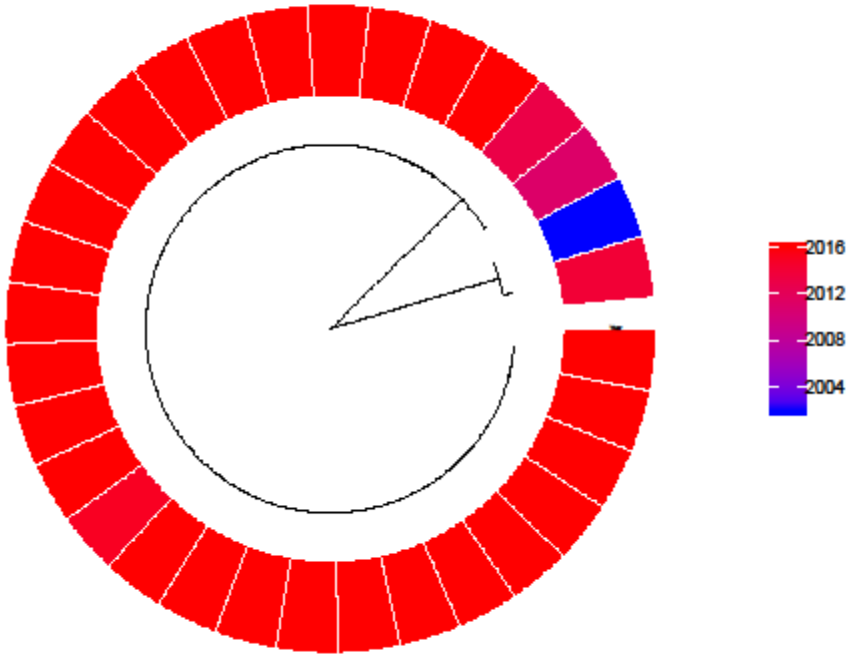
Supplementary Figure S11

Sub-lineages and outbreaks of L2 that were included in the analysis. Clades in colored in gray did not pass the DRT, clades colored in black passed the DRT. *: simple DRT passed, ** intermediate DRT passed, ***: stringent DRT passed. Dotted lines represent local two outbreaks from previous studies.



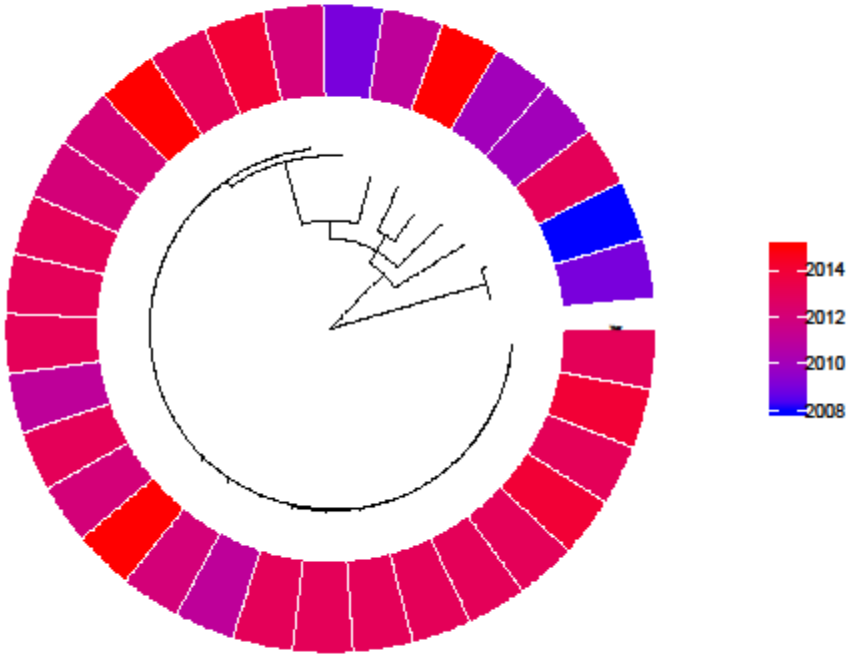
Supplementary Figure S12

Sub-lineages and outbreaks of L4 that were included in the analysis. Clades in colored in gray did not pass the DRT, clades colored in black passed the DRT. *: simple DRT passed, ** intermediate DRT passed, ***: stringent DRT passed. Dotted lines represent three outbreaks from previous studies.



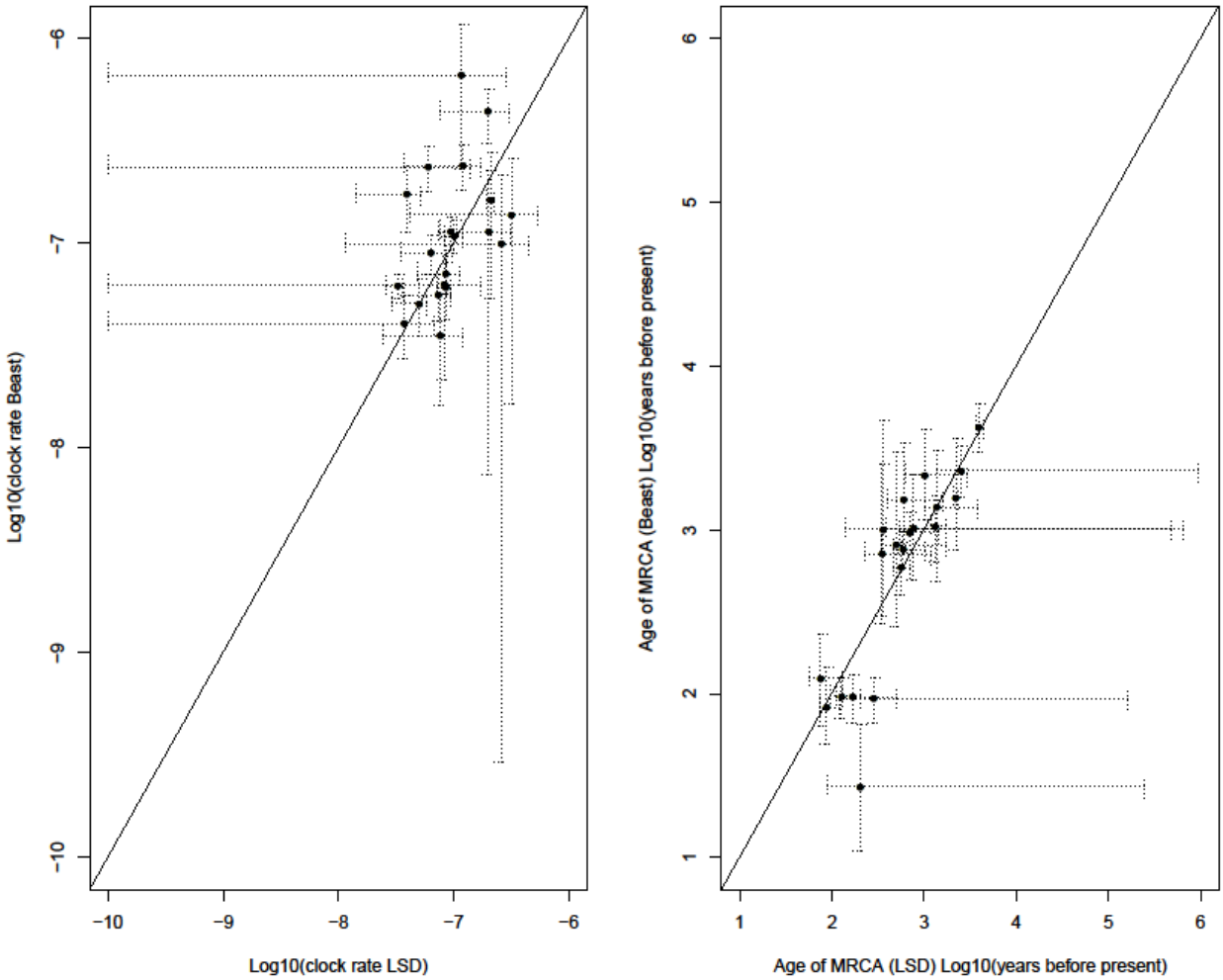
Supplementary Figure S13

Phylogenetic tree of data set L4_nc with tips colored according to the year of sampling



Supplementary Figure S14

Phylogenetic tree of data set Eldholm 2016 with tips colored according to the year of sampling



1274

1275 **Supplementary Figure S15**

1276 Comparison of results of LSD with Beast for data sets that passed the DRT, dotted lines represent the

1277 95% CI (LSD) or the 95% HPD (Beast)

1278

1279

1280

1281

Supplementary References

Crispell, J., Zadoks, R. N., Harris, S. R., Paterson, B., Collins, D. M., de-Lisle, G. W., ... & Kao, R. R. (2017). Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand. *BMC genomics*, *18*(1), 180.

Duchêne, S., Geoghegan, J. L., Holmes, E. C., & Ho, S. Y. (2016). Estimating evolutionary rates using time-structured data: a general comparison of phylogenetic methods. *Bioinformatics*, *32*(22), 3375-3379.

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., & Balloux, F. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature communications*, *6*, 7119.

Folkvardsen, D. B., Norman, A., Andersen, Å. B., Michael Rasmussen, E., Jelsbak, L., & Lillebaek, T. (2017). Genomic Epidemiology of a Major Mycobacterium tuberculosis Outbreak: Retrospective Cohort Study in a Low-Incidence Setting Using Sparse Time-Series Sampling. *The Journal of infectious diseases*, *216*(3), 366-374.

Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S. M., Rutaihwa, L. K., ... & Gagneux, S. (2018). Treemmer: a tool to reduce large phylogenetic data sets with minimal loss of diversity. *BMC bioinformatics*, *19*(1), 164.

Möller, S., du Plessis, L., & Stadler, T. (2018). Impact of the tree prior on estimating clock rates during epidemic outbreaks. *Proceedings of the National Academy of Sciences*, *115*(16), 4200-4205.

To, T. H., Jung, M., Lycett, S., & Gascuel, O. (2015). Fast dating using least-squares criteria and algorithms. *Systematic biology*, *65*(1), 82-97.