



HAL
open science

The paradoxes of *Mycobacterium tuberculosis* molecular evolution and consequences for the inference of tuberculosis emergence date

Rima Zein Eddine, F. Hak, A. Le Meur, C. Genestet, O. Dumitrescu, C. Guyeux, G. Senelle, C. Sola, G. Refrégier

► To cite this version:

Rima Zein Eddine, F. Hak, A. Le Meur, C. Genestet, O. Dumitrescu, et al.. The paradoxes of *Mycobacterium tuberculosis* molecular evolution and consequences for the inference of tuberculosis emergence date. *Tuberculosis*, 2023, 143, pp.102378. 10.1016/j.tube.2023.102378 . hal-04311241

HAL Id: hal-04311241

<https://hal.science/hal-04311241>

Submitted on 28 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Title : The paradoxes of *Mycobacterium tuberculosis* molecular evolution and consequences for the inference of tuberculosis emergence date

Authors

R. Zein Eddine¹, F. Hak², A. Le Meur², C. Genestet^{3,4}, O. Dumitrescu^{3,4}, C. Guyeux⁵, G. Senelle⁵, C. Sola^{6,7}, G. Refrégier^{2*}

Declarations of interest: none

Authors' agreement: all authors have approved the final version of the manuscript

Author contributions

Rima Zein-Eddine, Fiona Hak, Adrien Le Meur , Ccharlotte Genestet, Gaëtan Senelle performed data curation and formal analyses of data at the basis of this work. Guislaine Refrégier created the model. Christophe Guyeux, Oana Dumitrescu, Christophe Sola, Guislaine Refrégier participated in the funding acquisition. Guislaine Refrégier wrote the initial draft and performed the final editing. Christophe Guyeux, Adrien Le Meur, Christophe Sola participated in editing the manuscript. Adrien Le Meur created the figure.

1 **The paradoxes of *Mycobacterium tuberculosis* molecular evolution and consequences for** 2 **the inference of tuberculosis emergence date**

3 R. Zein Eddine¹, F. Hak², A. Le Meur², C. Genestet^{3,4}, O. Dumitrescu^{3,4}, C. Guyeux⁵, G. Senelle⁵, C. Sola^{6,7}, G.
4 Refrégier^{2*}

5 1 Laboratoire d'Optique et Biosciences, Ecole Polytechnique, Institut National de la Santé et de la Recherche
6 Médicale : U1182, Centre National de la Recherche Scientifique : UMR7645, France

7 2 Université Paris-Saclay, CNRS, AgroParisTech, Ecologie Systématique et Evolution, 91190, Gif-sur-Yvette,
8 France.

9 3 CIRI - Centre International de Recherche en Infectiologie, Ecole Normale Supérieure de
10 Lyon, Université Claude Bernard Lyon-1, Inserm U1111, CNRS UMR5308, Lyon, France

11 4 Hospices Civils de Lyon, Institut des Agents Infectieux, Laboratoire de bactériologie, Lyon, France³

12 5 DISC Computer Science Department, FEMTO-ST Institute, UMR 6174 CNRS, Univ. Bourgogne Franche-Comté
13 (UBFC), 16 Route de Gray, 25000 Besançon, France;

14 6 Université de Paris, IAME, UMR1137, INSERM, Paris, France and AP-HP, GHU Nord, service de
15 mycobactériologie spécialisée et de référence, Laboratoire associé du Centre National de Référence des
16 mycobactéries et résistance des mycobactéries aux antituberculeux (CNR-MyRMA), Paris, France

17 7 Université Paris-Saclay, Saint-Aubin, France

18 *corresponding author : guislaine.refregier@universite-paris-saclay.fr lab. Ecologie Systématique et Evolution,
19 12 rue 128, 91190, Gif-sur-Yvette, France. Tel : +33 1 69 15 77 20.

20 REVIEW

21 **Abstract**

22 The date of tuberculosis emergence has been the subject of such a long debate that it almost
23 resembles a Graal quest. New studies joining archaeological efforts with sequencing methods raise
24 high hopes for solving the question. Yet, inferring the emergence of the disease based on samples
25 from epidemics that occurred relatively long after this emergence requires a molecular clock for
26 the *Mycobacterium tuberculosis* complex. Several clocks inferred from different types of loci
27 and/or different samples, using both sound reasoning and reliable data, are actually very different,
28 which we call the paradoxes of *M. tuberculosis* molecular evolution.

29 After having presented these paradoxes, we will remind the limits of the molecular clocks used in
30 the different studies such as assumption of homogeneous substitution rate. We will then review
31 recent results that shed new light on the characteristics of *M. tuberculosis* molecular evolution:
32 traces of diverse selection pressures, the impact of host dynamics, etc. We provide some ideas on
33 what to do next to get nearer to the Graal. Among them, the collection of additional remains from
34 ancient tuberculosis seems still essential.

35 **Keywords:** mutation rate, substitution rate, positive selection, purifying selection, pathogen.

36 **Highlights:** Tuberculosis is still the first deadly infectious disease caused by a bacterium. It has
37 imposed a high death toll on humankind for millennia, but what key events underlie this success
38 is still not clear. More specifically, the date of its wide success, which we will refer to as tuberculosis
39 emergence, is still highly debated, between 70,000 to 6,000 years Before Present. We discuss the
40 relative reliability and the different assumptions at the basis of these datings and provide a model
41 of all parameters affecting pathogens molecular clocks. We argue that remains with ancient *M.*
42 *tuberculosis* are still needed to shed light on tuberculosis main diversification events.

43

44

45 **1. Introduction**

46 Identifying the role of microorganisms in infectious diseases was key to the increase in life expectancies
47 that has characterized the XXth century [1, 2]). Reconstructing the history of a specific disease helps
48 understanding future threats and thus avoid them or limit their impact. Phylogenetics and
49 phylogenomics infer the common ancestry and evolutionary history of organisms based on DNA
50 sequences, either from a selection of genes or from complete genomes [3]. It has proven to be a very
51 powerful tool to disentangle the origin of recent epidemics [4, 5]. On a wider timescale, the picture is
52 however less clear as exemplified by the debates on the precise origin of 1918 flu [6-8].

53 Tuberculosis is both an ancient disease and the presently most deadly bacteria-related infectious
54 disease. Works of art and bones from Predynastic Egypt and the Middle Kingdom testify to the
55 presence of the disease [9-11]. Tedious work from archaeologists continue to fuel the evidence of past
56 traces of tuberculosis [12, 13], suggesting a high prevalence of the disease already 10,000 y Before
57 Present (BP): Baker et al. identified osteological evidence highly suggestive of tuberculosis in Syria
58 around 10,000 y Before Present in 10% of the remains [12]. Yet the road seems still long to make a
59 clear history of tuberculosis disease using only osteological evidences of the disease. Interdisciplinary
60 studies mixing archaeology and phylogenomics emerge to try to provide definite answers on the global
61 history of tuberculosis [14], potentially reducing the interest in further archaeological explorations
62 with or without ancient DNA. We however want here to draw attention to the hypotheses underlying
63 these inferences. Here, we first present the different dates of tuberculosis emergence according to
64 various phylogenetical inferences, and we clarify the hypotheses of the underlying models. Second, we
65 will present compelling data that show how much these hypotheses are violated in the bacterium
66 evolution. We will focus on *M. tuberculosis* complex *sensu stricto* (hereafter referred to as MTBC) that
67 includes all human-adapted tuberculosis lineages L1-L9 and the animal strains. We exclude *M. canettii*

68 despite its high sequence identity with the rest of the lineages, due to its poor contribution to the
69 present human epidemics and widely different timescales [15].

70

71 **2. Various Inferences regarding the age of tuberculosis and the relying evidences**

72 Most phylogenomics studies infer divergence dates using a molecular clock in a Bayesian framework.
73 They assume that DNA evolves at the same pace along the whole phylogeny. This molecular clock is
74 usually assessed either from recent heterochronous sampling of recent epidemics, or from dated
75 fossils [16]. Pathogens codiverging with hosts (and vice-versa) benefit from a third possibility: relying
76 on the partner's phylogeny [17]. We detail below the different types of studies that tried to infer the
77 date of tuberculosis emergence.

78 The emergence date of tuberculosis was first estimated using synonymous mutations from four *M.*
79 *tuberculosis* genes and the application of a mutation rate derived from *E. coli* and *S. typhimurium* [18].
80 The use of other bacteria's rate seems reasonable as they evolve drug resistance in similar proportions
81 in lab experiments (David 1971). This study suggested that *M. tuberculosis* clonal expansion dates back
82 to 15,300-20,400 years Before Present (BP). However, a closer look at the data shows that this
83 inference derived from only four synonymous mutations from a sample of 31 isolates. In addition,
84 these rare mutations occurred in a heterogeneous subselection of 4 genes with very diverse functions
85 (housekeeping genes targeted by drugs, antigens) [18]. In addition, it is difficult to reconstruct how
86 well the selected isolates represented the global diversity. They might have corresponded only to L4
87 as it was the most widely sampled lineage at this time, or to modern MTBC (including L2 and potentially
88 L3 lineages), or to the whole diversity including L1 and the L5-L6-animal strains group. If the proposed
89 dating stands for L4 emergence, this would mean that global MTBC expansion would be twice older.
90 Kapur et al. dating might therefore only roughly reflect the actual date of tuberculosis emergence.

91 The second type of estimation derived from the mutation rate of minisatellites data [19]. This approach
92 has the advantage of circumventing potential biases due to low single nucleotide mutation rate. The
93 diversity of MIRU profiles leads to date the MTBC Most Recent Common Ancestor (MRCA) to 40,000
94 years BP. This dating corresponds to a coalescence time of main L4 sublineages such as LAM (L4.3) of
95 7,000 years BP [19].

96 Another type of estimations is based on a molecular clock derived from *M. tuberculosis* ancient DNA.
97 The corresponding studies took both synonymous and non-synonymous mutations without
98 distinction. This choice is motivated by the quest of precision in the mutation rate estimation. It is
99 supported by the clonal evolution of MTBC and an average ratio between synonymous and non-

100 synonymous mutations (dN/dS) close to 1 (~0.7) [20]. Such a ratio suggests that no selection occurred.
101 Yet, another possibility is the combination of positive/diversifying selection (dN/dS higher than 1) to
102 negative/purifying selection (dN/dS below 1) (see further discussion below). The first ancient genome
103 was sequenced from a Peruvian mummy infected by *M. pinnipedii* [21]. The inferred mutation rate was
104 estimated at $\sim 5 \cdot 10^{-8}$ subst/site/year. Applying this rate to the global phylogeny induces to date MTBC
105 MRCA at $\sim 6,000$ y before present [21]. Interestingly, new estimations based on L4 MTBC isolated from
106 Hungarian mummies come out with similar estimations (Brynildsrud et al, 2017; Sabin et al, 2020).

107 At last, another study inferring a molecular clock used the codivergence of *M. tuberculosis* with its
108 human hosts. The peculiar history and very restricted expansion of L4.2.2 sublineage allowed Refrégier
109 al (2016) to support the same molecular clock as Bos et al., dating MTBC MRCA to $\sim 6,000$ BP. Of note,
110 the authors do not argue in favour of this hypothesis as discussed below (Refrégier et al, 2016).
111 Interestingly, this article pinpoints the fact that the molecular clock inferred from heterochronous
112 sampling of recent epidemics is around twice faster as found by several independent studies at around
113 10^{-7} subst/site/year [22-24].

114 Apart from the studies using molecular clocks, Comas et al. dated MTBC clonal expansion using the
115 human phylogeny. Based on some similarities between the phylogenetic trees and geographic patterns
116 of *M. tuberculosis* and modern *H. sapiens* involved in the Out-of-Africa migration event, emergence
117 was reevaluated to $\sim 70,000$ y before Present, questioning the Neolithic nature of tuberculosis
118 emergence (Comas et al, 2013). The corresponding global mutation rate deriving from this inference
119 is $\sim 3 \cdot 10^{-9}$ subst/site/year which is around 1 order of magnitude less than that of *Y. pestis* and 3 orders
120 less than *S. pneumoniae*, and also 1 order of magnitude less than that of studies based on ancient DNA
121 [25].

122 Altogether, the estimated dates of tuberculosis emergence span a very wide time range, going from
123 6,000 to 70,000 years BP. Younger inferences were obtained using both synonymous and non-
124 synonymous mutations of ancient MTBC isolates, whereas inferences based solely on synonymous
125 mutations or minisatellites data or codivergence with the host come up with elder age of emergence.
126 We will next examine deeper why inferences based on both synonymous and non-synonymous
127 mutations might be erroneous.

128

129

130 **3. Assumptions at the basis of molecular dating**

131 Errors of several orders of magnitude can happen when inferring dates with sequences that do not
132 evolve with constant substitution rates *i.e.* with constant molecular clock [26]. Let us remind here that
133 substitution rate differs from the mutation rate by two characteristics: substitution looks at
134 accumulated modifications in the population after selection action, and substitution rate is usually
135 measured per time (per year) instead of per generation. Altogether, both the bacterial multiplication
136 rate, whether intra or interpatient, and selection are evolutionary forces controlling the impact of
137 mutation rate on the molecular clock (Grey ovals, Fig. 1). We will now detail how substitution rate (and
138 thus molecular clock) can vary among genes and how selection and bacterial multiplication may
139 influence it.

140 Many nucleotides in a genome do not evolve with the same substitution rate. This is obviously true
141 between genes when one has undergone horizontal gene transfer and thus changed cellular
142 environment that controls mutation rate, nucleotide concentrations, for all types of mutations [27].
143 The divergence between single nucleotide substitution rate is largely inflated between synonymous
144 and non-synonymous mutations as a consequence of selection: positive/diversifying selection is
145 characterized by the fixation of non-synonymous mutations conferring a selective advantage to the
146 organism, and negative/purifying selection corresponds to the elimination of non-synonymous
147 mutations that reduce organism fitness. This is actually the reason why the non-
148 synonymous/synonymous substitution ratio, the dN/dS, is indicative of different selection pressures
149 between genes [20]. Several studies have in fact documented the existence of positive/diversifying
150 selection pressures on MTBC genes. This includes genes targeted by drugs but also other genes
151 involved in host-pathogen interaction [28-31]. Recent exploration of trajectories of dN/dS showed
152 strikingly different dN/dS trajectories supporting a major past selection pressure on epitopes, and a
153 more recent pressure on toxin-antitoxin systems [32]. The selection on MTBC epitopes may relate to
154 the reciprocal selection of MTBC onto human genome [33]. Experimental evolution selecting for
155 biofilm formation also showed a high propensity for adaptation of MTBC, especially active upon
156 regulatory proteins, as well as an important role for epistatic interactions [34] (gene/sequences types
157 in red in Fig. 1). Moreover, mutation rate also depends on DNA repair systems and on oxidative damage
158 as recently shown by Moreno-Molina *et al.* [35] and Genestet *et al.*[31]. Last, duplicated genes such as
159 PE-PPE undergo occasional intragenomic recombination, which should also be taken into account if
160 their mutations are taken into account [36]. Obviously thus, the pace at which non-synonymous
161 mutations accumulate in different types of genes varies due to selection, potentially heterogeneous
162 mutation rate. This advocates against the common practice of merging synonymous and non-
163 synonymous mutations when building a molecular clock [37]. The current studies using ancient DNA
164 all suffer from this limitation.

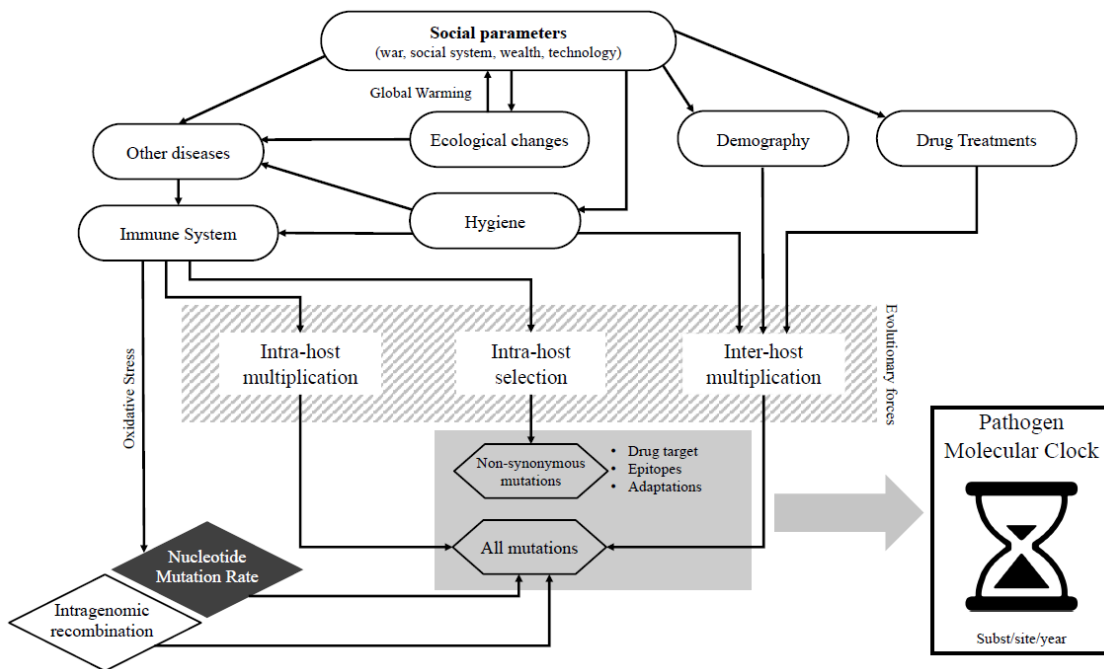
165 We showed that intra-host selection acting upon non-synonymous mutations impacts MTBC molecular
166 clock. We want to underline in addition that these selections are likely to have varied through time
167 and space. Indeed, intra-host selection depends on patient's drug treatment, but also on the
168 characteristics of the immune system (efficiency, adaptation). These characteristic first depend on the
169 host genetics [33]. They obviously also highly depend on other diseases. For instance, HIV-status is a
170 strong predictor of tuberculosis disease and tuberculosis regained strength with HIV dissemination
171 around the 80ies [38], but other diseases such as diabetes are also drivers of tuberculosis [39]. And of
172 course recently, COVID has shown to have indirectly impacted tuberculosis transmission by reducing
173 the diagnostic measures efficiency [40]. Other diseases may have greatly impacted the ability of the
174 immune system to fight tuberculosis over time and space, either because they exhaust the body,
175 reduce immune efficiency or because they may provide cross-immunity such as leprosy [41](Fig. 1).
176 Other key parameters for MTBC opportunities to disperse are: demography, which is impacted by
177 social parameters such as wars, economic wealth, etc and impacts migration; ecological changes linked
178 to present global warming or past cooling (the Little Ice Age) [42]

179 A last important point to bear in mind when considering molecular clock *versus* mutation rate is that
180 mutation rate is usually measured as a number of nucleotide changes per generation. This is because
181 most mutations are linked to replication infidelity and DNA repair [43-45]. But generations can have
182 different length time. Molecular clock may thus largely depend on the pace at which MTBC multiplies
183 inside the host (intrahost multiplication). Ford et al questioned this point by comparing the
184 accumulation of MTBC mutations in infected macaques carrying either latent or active tuberculosis
185 [46]. They could not detect a difference between the two categories but the power they had to do so
186 was limited due to the low number of replicates and the relatively short time of investigation, which
187 allowed to accumulate 0 to 3 substitutions in 91 to 507 days depending on the animal survival. In
188 addition, the study did not discriminate between synonymous and non-synonymous mutations.

189

190 We thus showed that non-synonymous mutations may have diverse substitution rates *i.e.* molecular
191 clocks, with one key evolutionary force at play being selection. This should prompt us to focus on
192 synonymous mutations. The synonymous mutation rate was explored shortly after the sequencing of
193 the second human MTBC strain, CDC1551 [47]. Interestingly, this rate appeared to vary highly between
194 genes. Recent deep exploration of diversity also suggests that various mutation rates prevail along the
195 genome of bacteria [48]. Synonymous mutations could thus be an incomplete solution to the problem
196 of identifying a reliable molecular clock.

197



199

200

Figure 1 – Parameters affecting molecular pathogens' clock, a model built for MTBC

201

We distinguish two types of targets: all mutations or only non-synonymous mutations (grey hexagons).

202

The main drivers of the difference between mutation and the molecular clock are intra and inter-host

203

multiplication and intra-host selection (evolutionary forces rectangle). Parameters affecting intra and

204

inter-host processes are indicated above. They include community parameters (social characteristics,

205

demography), global disease status (hygiene, other diseases prevalence) which impacts directly or

206

indirectly the immune status of the population and the opportunities for transmission *i.e.* inter-host

207

multiplication.

208

209

4. Conclusions

210

211

Since the model of epidemiologic transition, pathogens are thought to have shaped the history of

212

humankind (Caldwell 2001). *Mycobacterium tuberculosis*, with its incredibly low mutation rate, pushes

213

researchers working on phylogenetic inferences to merge data that clearly have a very diverse

214

dynamics due to selection, but also due to demography, the impact of other diseases and mutation

215

rate. Altogether, we have shown that no safe dating can be presently provided for tuberculosis

216

emergence. This holds for major diversification events of the agent of tuberculosis such as Lineage 4

217

expansion. As a reminder, we may speak for major diversification event for this lineage as 9 sublineages

218

have been consistently considered at an equivalent position in the taxonomy (L4.1 to L4.9). This major

219 diversification event could be due to Roman Empire expansion or Christian Crusades. In addition to the
220 emergence date of tuberculosis, this 1,000 y uncertainty on this major diversification event in MTBC is
221 also far from being solved.

222 Now that the reliability of sequencing is improving, that powerful methods exist and that the amount
223 of data is exploding, turning towards synonymous mutations and implementing in rich phylodynamics
224 model may help revisiting the possible history of tuberculosis. Yet, all assumptions underlying
225 phylodynamics, including host demography, MTBC mutation rate, selection, etc. should be kept in
226 mind to avoid exaggerated inferences. Independent confirmations are needed. They may come from
227 other diseases, human history, etc. Altogether, archaeological research, with more bones to gnaw on,
228 will definitively be of great help to strengthen or lower support of phylogenetic conclusions.

229

230 **Funding**

231 RZ has received funding from the European Union'S Horizon 2020 research and innovation programme
232 under the Marie Skłodowska-Curie grant agreement No 899987. ALM benefits from a PhD fellowship
233 from ANRS-MIE 2022-2, grant agreement n° 22476.

234

235 **Acknowledgements**

236 T. Giraud and E. Cambau are acknowledged for fruitful discussions on theoretical and practical grounds
237 related to the subject.

238

239 **Author contributions**

240 Rima Zein-Eddine, Fiona Hak, Adrien Le Meur , Ccharlotte Genestet, Gaëtan Senelle performed data
241 curation and formal analyses of data at the basis of this work. Guislaine Refrégier created the model.
242 Christophe Guyeux, Oana Dumitrescu, Christophe Sola, Guislaine Refrégier participated in the funding
243 acquisition. Guislaine Refrégier wrote the initial draft and performed the final editing. Christophe
244 Guyeux, Adrien Le Meur, Christophe Sola, Rima Zein-Eddine participated in editing the manuscript.
245 Adrien Le Meur created the figure.

246

247 **5. References**

248

- 249 1. Roser, M., E. Ortiz-Ospina, and H. Ritchie. *Life Expectancy*. 2019 [cited 2013 Jan. 2023];
 250 Available from: <https://ourworldindata.org/life-expectancy>.
- 251 2. Caldwell, J.C., *Population health in transition*. Bull World Health Organ, 2001. **79**(2): p. 159-
 252 60.
- 253 3. Rascovan, N., et al., *Emergence and Spread of Basal Lineages of Yersinia pestis during the*
 254 *Neolithic Decline*. Cell, 2019. **176**(1-2): p. 295-305 e10.
- 255 4. Weill, F.X., et al., *Genomic history of the seventh pandemic of cholera in Africa*. Science, 2017.
 256 **358**(6364): p. 785-789.
- 257 5. Piarroux, R., et al., *Understanding the cholera epidemic, Haiti*. Emerg Infect Dis, 2011. **17**(7):
 258 p. 1161-8.
- 259 6. Gibbs, M.J. and A.J. Gibbs, *Molecular virology: was the 1918 pandemic caused by a bird flu?*
 260 Nature, 2006. **440**(7088): p. E8; discussion E9-10.
- 261 7. Antonovics, J., M.E. Hood, and C.H. Baker, *Molecular virology: was the 1918 flu avian in*
 262 *origin?* Nature, 2006. **440**(7088): p. E9; discussion E9-10.
- 263 8. Taubenberger, J.K., *Influenza hemagglutinin attachment to target cells: 'birds do it, we do*
 264 *it...'* Future Virol, 2006. **1**(4): p. 415-418.
- 265 9. Crubezy, E., et al., *Identification of Mycobacterium DNA in an Egyptian Pott's disease of 5 400*
 266 *years old. Identificdtion dilDN de Mycobacterium dims un mul de POE &-yptien*
 267 *de 5400 ans*. C. R. Acad. Sci. Paris, 1997. **321**.
- 268 10. Zink, A.R., et al., *Characterization of Mycobacterium tuberculosis complex DNAs from*
 269 *Egyptian mummies by spoligotyping*. J Clin Microbiol, 2003. **41**(1): p. 359-67.
- 270 11. Wells, C., *Bones, bodies, and disease; evidence of disease and abnormality in early man*.
 271 Ancient peoples and places,. 1964, New York,: Praeger. 288 p.
- 272 12. Baker, O., et al., *Prehistory of human tuberculosis: Earliest evidence from the onset of animal*
 273 *husbandry in the Near East*.
 274 . Paléorient, 2017. **43**(2).
- 275 13. Spekker, O., et al., *A rare case of calvarial tuberculosis from the Avar Age (8th century CE)*
 276 *cemetery of Kaba-Bitozug (Hajdu-Bihar county, Hungary) - Pathogenesis and differential*
 277 *diagnostic aspects*. Tuberculosis (Edinb), 2022. **135**: p. 102226.
- 278 14. Sabin, S., et al., *A seventeenth-century Mycobacterium tuberculosis genome supports a*
 279 *Neolithic emergence of the Mycobacterium tuberculosis complex*. Genome Biol, 2020. **21**(1):
 280 p. 201.
- 281 15. Gutierrez, M.C., et al., *Ancient origin and gene mosaicism of the progenitor of*
 282 *Mycobacterium tuberculosis*. PLoS Pathog, 2005. **1**(1): p. e5.
- 283 16. Ho, S.Y., et al., *Time-dependent rates of molecular evolution*. Mol Ecol, 2011. **20**(15): p. 3087-
 284 101.
- 285 17. Wirth, T., A. Meyer, and M. Achtman, *Deciphering host migrations and origins by means of*
 286 *their microbes*. Mol Ecol, 2005. **14**(11): p. 3289-306.
- 287 18. Kapur, V., T.S. Whittam, and J.M. Musser, *Is Mycobacterium tuberculosis 15,000 years old?* J
 288 Infect Dis, 1994. **170**(5): p. 1348-9.
- 289 19. Wirth, T., et al., *Origin, spread and demography of the Mycobacterium tuberculosis complex*.
 290 PLoS Pathog, 2008. **4**(9): p. e1000160.
- 291 20. Yang, Z., et al., *Codon-substitution models for heterogeneous selection pressure at amino acid*
 292 *sites*. Genetics, 2000. **155**(1): p. 431-49.
- 293 21. Bos, K.I., et al., *Pre-Columbian mycobacterial genomes reveal seals as a source of New World*
 294 *human tuberculosis*. Nature, 2014.
- 295 22. Bryant, J.M., et al., *Inferring patient to patient transmission of Mycobacterium tuberculosis*
 296 *from whole genome sequencing data*. BMC Infect Dis, 2013. **13**: p. 110.

- 297 23. Roetzer, A., et al., *Whole genome sequencing versus traditional genotyping for investigation*
298 *of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study.*
299 PLoS Med, 2013. **10**(2): p. e1001387.
- 300 24. Walker, T.M., et al., *Whole-genome sequencing to delineate Mycobacterium tuberculosis*
301 *outbreaks: a retrospective observational study.* Lancet Infect Dis, 2013. **13**(2): p. 137-46.
- 302 25. Comas, I., et al., *Out-of-Africa migration and Neolithic coexpansion of Mycobacterium*
303 *tuberculosis with modern humans.* Nat Genet, 2013. **45**(10): p. 1176-82.
- 304 26. Kuo, C.H. and H. Ochman, *Inferring clocks when lacking rocks: the variable rates of molecular*
305 *evolution in bacteria.* Biol Direct, 2009. **4**: p. 35.
- 306 27. Vos, M., et al., *Rates of Lateral Gene Transfer in Prokaryotes: High but Why?* Trends
307 Microbiol, 2015. **23**(10): p. 598-605.
- 308 28. Chiner-Oms, A. and I. Comas, *Large genomics datasets shed light on the evolution of the*
309 *Mycobacterium tuberculosis complex.* Infect Genet Evol, 2019. **72**: p. 10-15.
- 310 29. Osorio, N.S., et al., *Evidence for diversifying selection in a set of Mycobacterium tuberculosis*
311 *genes in response to antibiotic- and nonantibiotic-related pressure.* Mol Biol Evol, 2013.
312 **30**(6): p. 1326-36.
- 313 30. Tantivitayakul, P., et al., *Homoplastic single nucleotide polymorphisms contributed to*
314 *phenotypic diversity in Mycobacterium tuberculosis.* Sci Rep, 2020. **10**(1): p. 8024.
- 315 31. Genestet, C., et al., *Mycobacterium tuberculosis genetic features associated with pulmonary*
316 *tuberculosis severity.* Int J Infect Dis, 2022. **125**: p. 74-83.
- 317 32. Chiner-Oms, A., et al., *Gene evolutionary trajectories in Mycobacterium tuberculosis reveal*
318 *temporal signs of selection.* Proc Natl Acad Sci U S A, 2022. **119**(17): p. e2113600119.
- 319 33. Kerner, G., et al., *Human ancient DNA analyses reveal the high burden of tuberculosis in*
320 *Europeans over the last 2,000 years.* Am J Hum Genet, 2021. **108**(3): p. 517-524.
- 321 34. Smith, T.M., et al., *Rapid adaptation of a complex trait during experimental evolution of*
322 *Mycobacterium tuberculosis.* Elife, 2022. **11**.
- 323 35. Moreno-Molina, M., et al., *Genomic analyses of Mycobacterium tuberculosis from human*
324 *lung resections reveal a high frequency of polyclonal infections.* Nat Commun, 2021. **12**(1): p.
325 2716.
- 326 36. Liu, X., et al., *Evidence for recombination in Mycobacterium tuberculosis.* J Bacteriol, 2006.
327 **188**(23): p. 8169-77.
- 328 37. Menardo, F., et al., *The molecular clock of Mycobacterium tuberculosis.* PLoS Pathog, 2019.
329 **15**(9): p. e1008067.
- 330 38. WHO, *Global Tuberculosis Control : Surveillance, Planning, Financing.* 2004, WHO, Geneva,
331 Switzerland.
- 332 39. Mathema, B., et al., *Drivers of Tuberculosis Transmission.* J Infect Dis, 2017. **216**(suppl_6): p.
333 S644-S653.
- 334 40. WHO, *Global tuberculosis report 2022.* 2022, World Health Organization: Geneva. p. 1-68.
- 335 41. Hambridge, T., et al., *Mycobacterium leprae transmission characteristics during the declining*
336 *stages of leprosy incidence: A systematic review.* PLoS Negl Trop Dis, 2021. **15**(5): p.
337 e0009436.
- 338 42. Menardo, F., *Understanding drivers of phylogenetic clustering and terminal branch lengths*
339 *distribution in epidemics of Mycobacterium tuberculosis.* Elife, 2022. **11**.
- 340 43. Castaneda-Garcia, A., et al., *A non-canonical mismatch repair pathway in prokaryotes.* Nat
341 Commun, 2017. **8**: p. 14246.
- 342 44. Kunkel, T.A. and K. Bebenek, *DNA replication fidelity.* Annu Rev Biochem, 2000. **69**: p. 497-
343 529.
- 344 45. Rock, J.M., et al., *DNA replication fidelity in Mycobacterium tuberculosis is mediated by an*
345 *ancestral prokaryotic proofreader.* Nat Genet, 2015. **47**(6): p. 677-81.
- 346 46. Ford, C.B., et al., *Use of whole genome sequencing to estimate the mutation rate of*
347 *Mycobacterium tuberculosis during latent infection.* Nat Genet, 2011. **43**(5): p. 482-6.

- 348 47. Hughes, A.L., R. Friedman, and M. Murray, *Genomewide Pattern of Synonymous Nucleotide*
349 *Substitution in Two Complete Genomes of Mycobacterium tuberculosis*. *Emerg Infect Dis*,
350 2002. **8**(11): p. 1342-6.
- 351 48. Jee, J., et al., *Rates and mechanisms of bacterial mutagenesis from maximum-depth*
352 *sequencing*. *Nature*, 2016. **534**(7609): p. 693-6.
- 353

