

1 A seventeenth-century *Mycobacterium tuberculosis* genome supports a Neolithic emergence
2 of the *Mycobacterium tuberculosis* complex.

3

4 Susanna Sabin¹ (sabin@shh.mpg.de), Alexander Herbig¹ (herbig@shh.mpg.de), Åshild J.

5 Vågane¹ (vagane@shh.mpg.de), Torbjörn Ahlström² (torbjorn.ahlstrom@ark.lu.se), Gracijela

6 Bozovic³ (gracijela.bozovic@med.lu.se), Caroline Arcini⁴

7 (caroline.ahlstrom.arcini@arkeologerna.com), Denise Kühnert^{5*} (kuehnert@shh.mpg.de),

8 Kirsten I. Bos^{1*} (bos@shh.mpg.de)

9 1. Department of Archaeogenetics, Max Planck Institute for the Science of Human
10 History, Jena, Germany 07745

11 2. Department of Archaeology and Ancient History, Lund University, Lund, Sweden 221
12 00

13 3. Department of Medical Imaging and Clinical Physiology, Skåne University Hospital
14 Lund and Lund University, Lund, Sweden 221 00

15 4. Arkeologerna, National Historical Museum, Lund, Sweden 226 60

16 5. Transmission, Infection, Diversification & Evolution Group, Max Planck Institute for
17 the Science of Human History, Jena, Germany 07745

18 *Corresponding authors

19

20 Running title: A seventeenth-century *Mycobacterium tuberculosis* genome supports a
21 Neolithic emergence of the *Mycobacterium tuberculosis* complex

22 Keywords: tuberculosis, ancient DNA, *Mycobacterium tuberculosis*, molecular dating,
23 metagenomics

24 ABSTRACT

25 Background:

26 Although tuberculosis accounts for the highest mortality from a bacterial infection on a global
27 scale, questions persist regarding its origin. One hypothesis based on modern
28 *Mycobacterium tuberculosis* complex (MTBC) genomes suggests their most recent common
29 ancestor (MRCA) followed human migrations out of Africa ~70,000 years before present
30 (BP). However, studies using ancient genomes as calibration points have yielded much
31 younger MRCA dates of less than 6,000 years. Here we aim to address this discrepancy
32 through the analysis of the highest-coverage and highest quality ancient MTBC genome
33 available to date, reconstructed from a calcified lung nodule of Bishop Peder Winstrup of
34 Lund (b. 1605 – d. 1697).

35 Results:

36 A metagenomic approach for taxonomic classification of whole DNA content permitted the
37 identification of abundant DNA belonging to the human host and the MTBC, with few non-TB
38 bacterial taxa comprising the background. Subsequent genomic enrichment enabled the
39 reconstruction of a 141-fold coverage *M. tuberculosis* genome. In utilizing this high-quality,
40 high-coverage 17th century *M. tuberculosis* genome as a calibration point for dating the
41 MTBC, we employed multiple Bayesian tree models, including birth-death models, which
42 allowed us to model pathogen population dynamics and data sampling strategies more
43 realistically than those based on the coalescent.

44 Conclusions

45 The results of our metagenomic analysis demonstrate the unique preservation environment
46 calcified nodules provide for DNA. Importantly, we estimate an MRCA date for the MTBC of
47 3683 BP (2253-5821 BP) and for Lineage 4 of 1651 BP (946-2575 BP) using multiple
48 models, confirming a Neolithic emergence for the MTBC.

49 BACKGROUND

50 Tuberculosis, caused by organisms in the *Mycobacterium tuberculosis* complex
51 (MTBC), has taken on renewed relevance and urgency in the 21st century due to its global

52 distribution, its high morbidity, and the rise of antibiotic resistant strains (1). The difficulty in
53 disease management and treatment, combined with the massive reservoir the pathogen
54 maintains in human populations through latent infection (2), makes tuberculosis a pressing
55 public health challenge. Despite this, controversy exists regarding the history of the
56 relationship between members of the MTBC and their human hosts.

57 Existing literature suggests two most recent common ancestor (MRCA) dates for the
58 MTBC based on the application of Bayesian molecular dating to genome-wide
59 *Mycobacterium tuberculosis* data. One estimate suggests the extant MTBC emerged
60 through a bottleneck approximately 70,000 years ago, coincident with major migrations of
61 humans out of Africa (3). This estimate was reached using exclusively modern *M.*
62 *tuberculosis* genomes, with internal nodes of the MTBC calibrated by extrapolated dates for
63 major human migrations (3). This estimate relied on congruence between the topology of
64 MTBC and human mitochondrial phylogenies, but this congruence does not extend to
65 human Y chromosome phylogeographic structure (4). As an alternative approach, the first
66 publication of ancient MTBC genomes utilized radiocarbon dates as direct calibration points
67 to infer mutation rates, and yielded an MRCA date for the complex of less than 6,000 years
68 (5). This younger emergence was later supported by mutation rates estimated within the
69 pervasive Lineage 4 (L4) of the MTBC, using four *M. tuberculosis* genomes from the late 18th
70 and early 19th centuries (6).

71 Despite the agreement in studies that have relied on ancient DNA calibration so far,
72 dating of the MTBC emergence remains controversial. Such a young age cannot account for
73 purported detection of MTBC DNA in archaeological material that predates the MRCA
74 estimate (e.g. Baker et al. 2015; Hershkovitz et al. 2008; Masson et al. 2013; Rothschild et
75 al. 2001), the authenticity of which has been challenged (11). Furthermore, constancy in
76 mutation rates of the MTBC has been challenged on account of observed rate variation in
77 modern lineages, combined with the unquantified effects of latency (12). The ancient
78 genomes presented by Bos and colleagues, though isolated from human remains, were
79 most closely related to *Mycobacterium pinnipedii*, a lineage of the MTBC associated with

80 infections in seals and sea lions today (5). Given our unfamiliarity with the demographic
81 history of tuberculosis in sea mammal populations (13), identical substitution rates between
82 the pinniped lineage and human-adapted lineages of the MTBC cannot be assumed.
83 Additionally, the identification of true genetic changes in archaeological specimens can be
84 difficult given the similarities between MTBC and environmental mycobacterial DNA from the
85 depositional context (14). Though the ancient genomes published by Kay and colleagues
86 belonged to human-adapted lineages of the MTBC, and the confounding environmental
87 signals were significantly reduced by their funerary context in crypts, two of the four
88 genomes used for molecular dating were derived from mixed-strain infections (6). By
89 necessity, diversity derived in each genome would have to be ignored for them to be
90 computationally distinguished (6). Though ancient DNA is a valuable tool for answering the
91 question of when the MTBC emerged, the available ancient data remains sparse and subject
92 to case-by-case challenges.

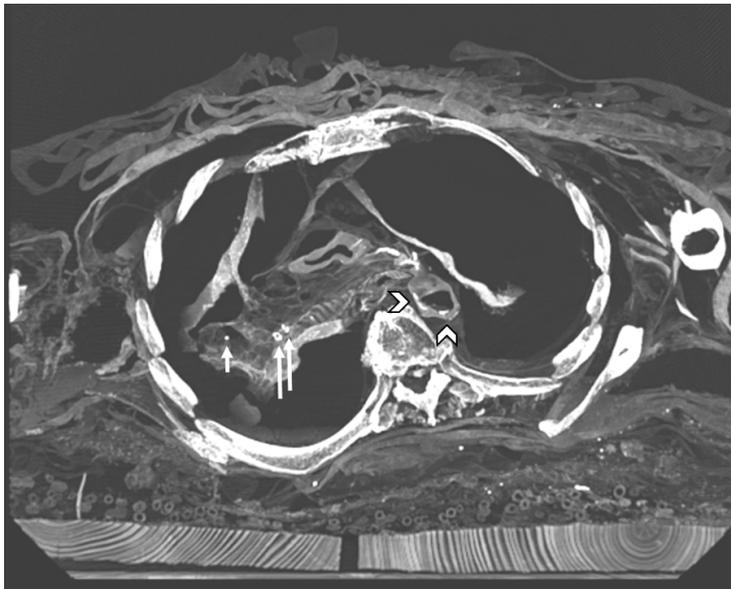
93 Here, we contribute to clarifying the timing of the emergence of the MTBC and L4
94 using multiple Bayesian models of varying complexity through the analysis of a high-
95 coverage 17th century *M. tuberculosis* genome extracted from a calcified lung nodule.
96 Removed from naturally mummified remains, the nodule provided an excellent preservation
97 environment for the pathogen, exhibiting minimal infiltration by exogenous bacteria. The
98 nodule and surrounding lung tissue also showed exceptional preservation of host DNA, thus
99 showing promise for this tissue type in ancient DNA investigations.

100 RESULTS

101 **Pathogen identification**

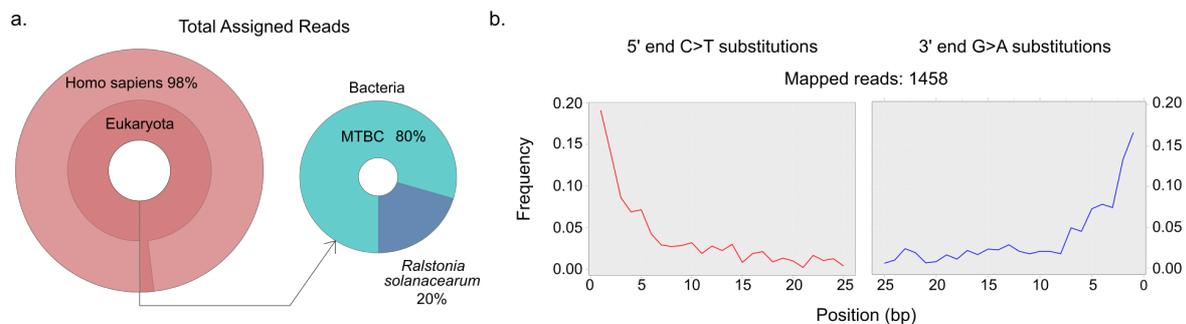
102 Computed tomography (CT) scans of the mummified remains of Bishop Peder
103 Winstrup of Lund revealed a calcified granuloma a few millimeters (mm) in size in the
104 collapsed right lung together with two ~5 mm calcifications in the right hilum (Figure 1).
105 Primary tuberculosis causes parenchymal changes and ipsilateral hilar lymphadenopathy
106 that is more common on the right side (15). Upon resolution it can leave a parenchymal scar,
107 a small calcified granuloma (Ghon Focus), and calcified hilar nodes, which are together

108 called a Ranke complex. In imaging this complex is suggestive of previous tuberculosis
109 infection, although histoplasmosis can have the same appearance (16). Histoplasmosis,
110 however, is very rare in Scandinavia and more often seen in other parts of the world (e.g.
111 the Americas) (17). The imaging findings were therefore considered to result from previous
112 primary tuberculosis. One of the calcified hilar nodes was extracted from the remains during
113 video-assisted thoracoscopic surgery, guided by fluoroscopy. The extracted material was
114 further subsampled for genetic analysis. DNA was extracted from the nodule and
115 accompanying lung tissue using protocols optimized for the recovery of ancient, chemically
116 degraded, fragmentary genetic material (18). The metagenomic library was shotgun
117 sequenced to a depth of approximately 3.7 million reads.



118
119 **Figure 1. CT image of Ranke complex.** CT image of Peder Winstrup's chest in a slightly
120 angled axial plane with the short arrow showing a small calcified granuloma in the probable
121 upper lobe of the collapsed right lung, and two approximately 5 mm calcifications in the right
122 hilum together suggesting a Ranke complex and previous primary tuberculosis. The more
123 lateral of the two hilar calcifications was extracted for further analysis. In addition, there are
124 calcifications in the descending aorta proposing atherosclerosis (arrowhead).
125

126 Adapter-clipped and base quality filtered reads were taxonomically binned with MALT
127 (19) against the full NCBI Nucleotide database ('nt', April 2016). In this process, 3,515,715
128 reads, or 95% of the metagenomic reads, could be assigned to taxa contained within the
129 database. Visual analysis of the metagenomic profile in MEGAN6 (20) revealed the majority
130 of these reads, 2,833,403 or 81%, were assigned to *Homo sapiens*. A further 1,724 reads
131 assigned to the *Mycobacterium tuberculosis* complex (MTBC) node. Importantly, no other
132 taxa in the genus *Mycobacterium* were identified, and the only other identified bacterial
133 taxon was *Ralstonia solanacearum* (Figure 2a), a soil-dwelling plant pathogen frequently
134 identified in metagenomic profiles of archaeological samples (21,22) (Table S1 in Additional
135 File 1).



136
137 **Figure 2. Screening of sequencing data from LUND1 shows preservation of host and**
138 **pathogen DNA.** A) Krona plots reflecting the metagenomic composition of the lung nodule.
139 The majority of sequencing reads were aligned to *Homo sapiens* (n=2,833,403),
140 demonstrating extensive preservation of host DNA. A small portion of reads aligned to
141 bacterial organisms, and 80% of these reads were assigned to the MTBC node (n=1,724).
142 B) Damage plots generated from sequencing reads mapped directly to a reconstructed
143 MTBC ancestor genome (23), demonstrating a pattern characteristic of ancient DNA.

144 Pre-processed reads were mapped to both the hg19 human reference genome and a
145 reconstructed MTBC ancestor (TB ancestor) (23) using BWA as implemented in the Efficient
146 Ancient Genome Reconstruction (EAGER) pipeline (24). Reads aligned to hg19 with direct
147 mapping constituted an impressive 88% of the total sequencing data (Table S2 in Additional
148 File 1). Human mitochondrial contamination was extremely low, estimated at only 1-3%

149 using Schmutzi (25) (Additional File 2). Reads were also mapped to the TB ancestor (Table
150 1). After map quality filtering and read de-duplication, 1,458 reads, or 0.045% of the total
151 sequencing data, aligned to the reference (Table 1), and exhibited cytosine-to-thymine
152 damage patterns indicative of authentic ancient DNA (Figure 2b) (26,27). Qualitative
153 preservation of the tuberculosis DNA was slightly better than that of the human DNA, as the
154 damage was greater in the latter (Table S2 in Additional File 1). Laboratory-based
155 contamination, as monitored by negative controls during the extraction and library
156 preparation processes, could be ruled out as the source of this DNA (Table S3 in Additional
157 File 1).

158 **Genomic enrichment and reconstruction**

159 Due to the clear but low-abundance MTBC signal, a uracil DNA glycosylase (UDG)
160 library was constructed to remove DNA lesions caused by hydrolytic deamination of cytosine
161 residues (28) and enriched with an in-solution capture (29,30) designed to target genome-
162 wide data representing the full diversity of the MTBC (see METHODS). The capture probes
163 are based on the TB ancestor genome (23), which is equidistant from all lineages of the
164 MTBC. The enriched library was sequenced using a paired-end, 150-cycle Illumina
165 sequencing kit to obtain a full fragment-length distribution (Figure S1 in Additional File 2).
166 The resulting sequencing data was then aligned to the hypothetical TB ancestor genome
167 (23), and the mapping statistics were compared with those from the screening data to
168 assess enrichment (Table 1). Enrichment increased the proportion of endogenous MTBC
169 DNA content by three orders of magnitude, from 0.045% to 45.652%, and deep sequencing
170 yielded genome-wide data at an average coverage of approximately 141.5-fold. The mapped
171 reads have an average fragment length of ~66 base pairs (Table 1).

172

Pre/post	Library treatment	Processed reads pre-mapping (<i>n</i>)	Unique mapped reads, quality-filtered (<i>n</i>)	Endogenous DNA (%)	Mean fold coverage	Mean fragment length (bp)	GC content (%)
Pre-capture	non-UDG	3696712	1458	0.045	0.018	54.31	63.89
Post-capture	UDG	59091507	9482901	45.652	141.5062	65.83	62.96

173 **Table 1. Mapping statistics for LUND1 libraries.** A comparison of the mapping statistics
174 for the non-UDG screening library and UDG-treated MTBC enriched library of LUND1 when
175 aligned to the MTBC ancestor genome (23). For full EAGER output, see Table S2 in
176 Additional File 1.

177

178 We further evaluated the quality of the reconstructed genome by quantifying the
179 amount of heterozygous positions (see METHODS). Derived alleles represented by 10-90%
180 of the reads covering a given position with five or more reads of coverage were counted.
181 Only 24 heterozygous sites were counted across all variant positions in LUND1. As a
182 comparison, the other high-coverage (~125 fold) ancient genome included here – body92
183 from Kay et al. 2015 – contained 70 heterozygous positions.

184 **Phylogeny and dating**

185 Preliminary phylogenetic analysis using neighbor joining (Figures S2 and S3 in
186 Additional File 2), maximum likelihood (Figures S4 and S5 in Additional File 2), and
187 maximum parsimony trees (Figures S6 and S7 in Additional File 2) indicated that LUND1
188 groups within the L4 strain diversity of the MTBC, and more specifically, within the
189 L4.10/PGG3 sublineage. This sublineage was recently defined by Stucki and colleagues as
190 the clade containing L4.7, L4.8, and L4.9 (31) according to the widely-accepted Coll
191 nomenclature (32). Following this, we constructed two datasets to support molecular dating
192 of the full MTBC (Table S4 in Additional File 1) and L4 of the MTBC (Table S5 in Additional
193 File 1).

194 The dataset reflecting extant diversity of the MTBC was compiled as reported
195 elsewhere (5), with six ancient genomes as calibration points. These included LUND1; two
196 additional ancient genomes, body80 and body92, extracted from late 18th and early 19th
197 century Hungarian mummies (6); and three human-isolated *Mycobacterium pinnipedii* strains
198 from Peru (5), encompassing all available ancient *M. tuberculosis* genomes with sufficient
199 coverage to call SNPs confidently after stringent mapping with BWA (33) (see METHODS;
200 Table S4 in Additional File 1). *Mycobacterium canettii* was used as an outgroup. In

201 generating an alignment of variant positions in this dataset, we excluded repetitive regions
202 and regions at risk of cross-mapping with other organisms as done previously (5), as well as
203 potentially imported sites from recombination events, which were identified using
204 ClonalFrameML (34) (Table S6 in Additional File 1). We chose to exclude these potential
205 recombination events despite *M. tuberculosis* being generally recognized as a largely clonal
206 organism with minimal recombination or horizontal gene transfer, as this is still a point of
207 contention (35). Only twenty-three variant sites were lost from the full MTBC alignment as
208 potential imports. We called a total of 42,856 variable positions in the dataset as aligned to
209 the TB ancestor genome. After incompletely represented sites were excluded, 11,716 were
210 carried forward for downstream analysis.

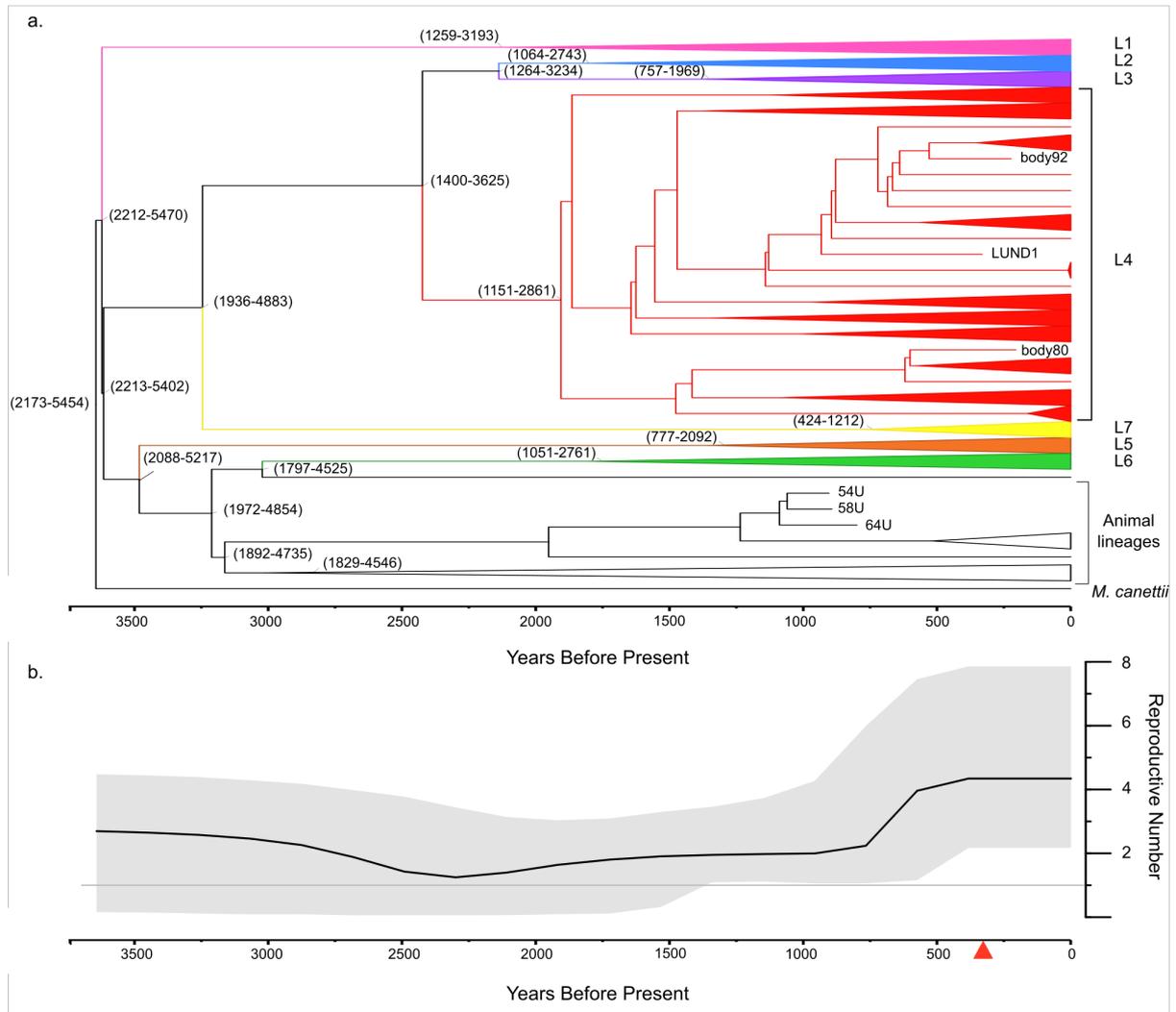
211 To explore the impact of the selected tree prior and clock model, we ran multiple
212 variations of models as available for use in BEAST2 (36). We first used both a strict and a
213 relaxed clock model together with a constant coalescent model (CC+strict, CC+UCLD). We
214 found there to be minimal difference between the inferred rates estimated by the two
215 models. This finding, in addition to the low rate variance estimated in all models, suggests
216 there is little rate variation between known branches of the MTBC. Nevertheless, the relaxed
217 clock appeared to have a slightly better performance (Table 2). To experiment with models
218 that allowed for dynamic populations, we applied a Bayesian skyline (SKY+UCLD) and birth-
219 death skyline prior (BDSKY+UCLD) combined with a relaxed clock model. In the
220 BDSKY+UCLD model, the tree was conditioned on the root. To our knowledge, this is the
221 first instance of a birth-death tree prior being used to infer evolutionary dynamics of the
222 MTBC while using ancient data for tip calibration.

223 A calibrated maximum clade credibility (MCC) tree was generated for the
224 BDSKY+UCLD model, with 3683 years before present (BP) (95% highest posterior density
225 [95% HPD] interval: 2253 – 5821 BP) as an estimated date of emergence for the MTBC
226 (Figure 3a). Tree topology agrees with previously presented phylogenetic analyses of the full
227 MTBC (3,5,37). A birth-death skyline plot illustrates the flux in the effective reproduction
228 number (R) over time (Figure 3b). In an outbreak setting, R refers to the average number of

229 secondary cases stemming from a single infection, and an epidemic event is inferred when
 230 the value is greater than one. However, for the data at hand, $R > 1$ translates to lineage
 231 diversification rates exceeding lineage death/extinction. Since there is no data representing
 232 the period between ~1000 years ago and the emergence of the MTBC, there is much
 233 uncertainty in the related estimates. From around 1300 BP the 95% HPD excludes 1,
 234 indicating a positive net diversification rate, with a significant increase between 974 and 390
 235 BP (odds ratio=10.00054).

Model	Mean Likelihood	Mean Rate (95% HPD)	Mean Rate Variance (95% HPD)	Mean Tree Height (95% HPD)
BDSKY+UCLD	-6123180.475	1.303E-8 (6.9753E-9, 1.8348E-8)	1.3656E-17 (2.4838E-18, 2.5884E-17)	3683.203 (2253.2836, 5820.8405)
CC+UCLD	-6123187.492	1.214E-8 (7.1934E-9, 1.6448E-8)	1.2459E-17 (2.833E-18, 2.3969E-17)	4172.1961 (2585.2349, 6119.744)
SKY+UCLD	-6123279.053	1.3294E-8 (8.9335E-9, 1.7461E-8)	1.4147E-17 (5.1837E-18, 2.4356E-17)	3540.7193 (2453.8322, 4829.7259)
CC+strict	-6123688.933	1.1573E-8 (8.6397E-9, 1.4509E-8)	NA	4453.1162 (3330.1516, 5619.3974)

236 **Table 2. Model comparison for full MTBC dataset.** Parameter estimates from four models
 237 applied to the full MTBC dataset: constant coalescent with uncorrelated lognormal clock
 238 (CC+UCLD), constant coalescent with strict clock (CC+strict), Bayesian skyline coalescent
 239 with uncorrelated lognormal clock (SKY+UCLD), and birth-death skyline with uncorrelated
 240 lognormal clock (BDSKY+UCLD).



241

242 **Figure 3. MTBC maximum clade credibility tree and birth-death skyline plot.** A) This
 243 MCC tree of mean heights was generated from the BDSKY+UCLD model as applied to the
 244 full MTBC dataset. Modern genomes are collapsed according to lineage (labeled on the right
 245 side). The ancient genomes are labeled with their sample name. The outgroup is labeled as
 246 “*M. canettii*.” The 95% HPD intervals of selected node heights are indicated as (lower
 247 boundary - upper boundary) in years before present. The time scale is expressed as years
 248 before present, with the most recent time as 2010. B) The black line indicates median
 249 reproductive number over time (reproductive number set to 5 dimensions, see Additional File
 250 2). The shaded grey area represents the 95% HPD interval of the reproductive number. The
 251 grey line indicates a reproductive number of 1. The red triangle on the timeline indicates the
 252 temporal position of LUND1.

253

254 The L4 dataset includes LUND1 and the two Hungarian mummies described above
255 (6) as calibration points. We selected 149 modern genomes representative of the known
256 diversity of L4 from previously published datasets (Additional File 2) (3,23,31). A modern
257 Lineage 2 (L2) genome was used as an outgroup. After the exclusion of sites as discussed
258 above (Table S7 in Additional File 1), a SNP alignment of these genomes in reference to the
259 reconstructed TB ancestor genome (23) included a total of 17,333 variant positions,
260 excluding positions unique to the L2 outgroup. Only fifteen variant sites were lost from the L4
261 dataset alignment. After sites missing from any alignment in the dataset were excluded from
262 downstream analysis, 10,009 SNPs remained for phylogenetic inference. A total of 810
263 SNPs were identified in LUND1, of which 126 were unique to this genome. A SNP effect
264 analysis (38) was subsequently performed on these derived positions (Additional File 2;
265 Table S8 in Additional File 1).

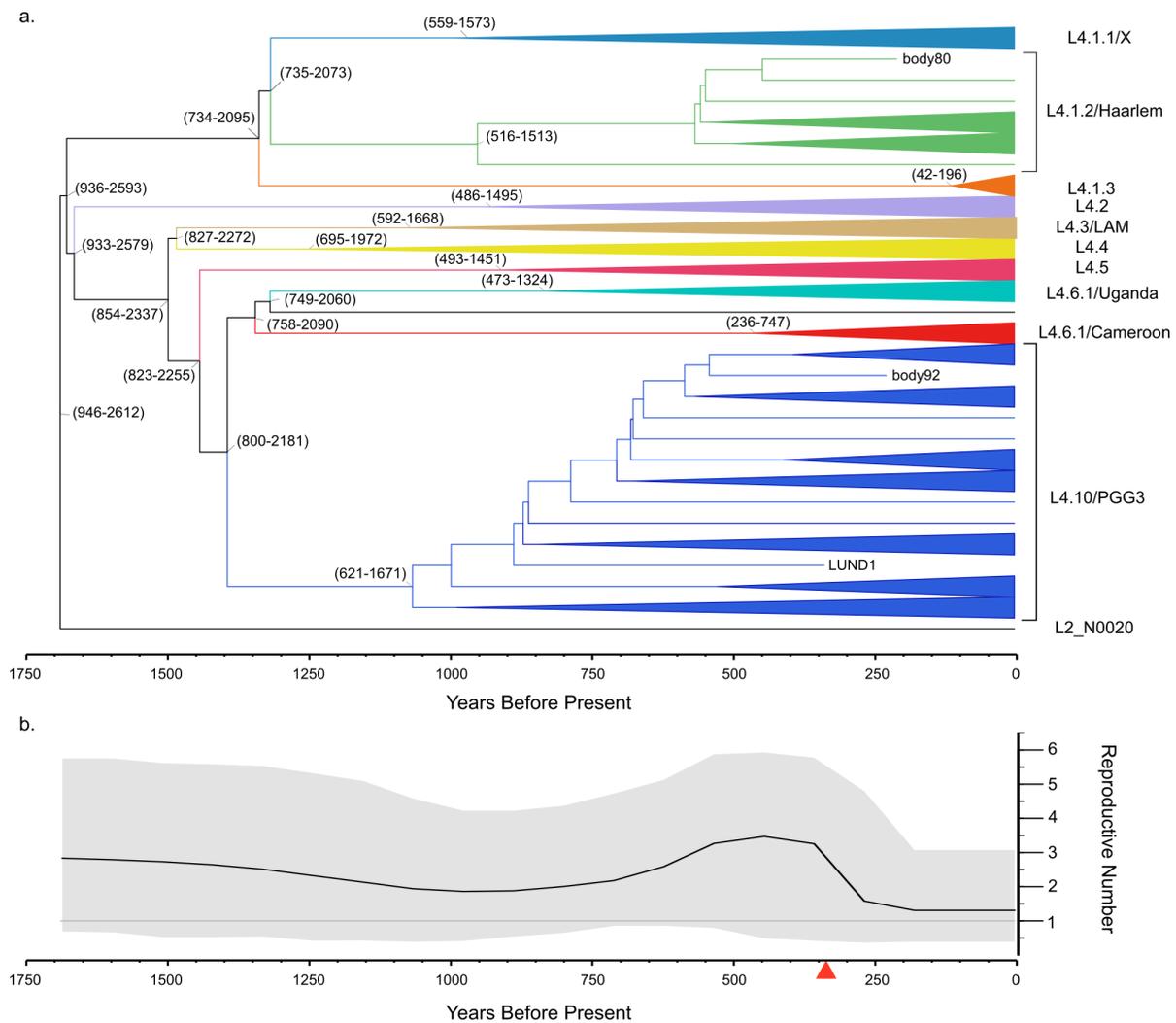
266 We applied the same models as described above for the full MTBC dataset, with the
267 addition of a birth-death skyline model conditioned on the origin of the root
268 (BDSKY+UCLD+origin). All mean tree heights are within 250 years of each other and the
269 95% HPD intervals largely overlap. As an informal model comparison, the BDSKY+UCLD
270 model shows the highest marginal likelihood values. We employed the
271 BDSKY+UCLD+origin model to determine if the estimated origin of the L4 dataset agreed
272 with the tree height estimates for the full MTBC dataset. Intriguingly, the estimated origin
273 parameter (Table 3), or the ancestor of the tree root, largely overlaps with the 95% HPD
274 range for MTBC tree height as seen in Table 2.

275 A calibrated MCC tree (Figure 4a) was generated based on the BDSKY+UCLD
276 model for the L4 dataset. This model yielded an estimated date of emergence for L4 of 1650
277 BP (95% HPD: 946-2575 BP). The tree reflects the ten-sublineage topology presented by
278 Stucki and colleagues (31), with LUND1 grouping with the L4.10/PGG3 sublineage. A birth-
279 death skyline plot was also generated (Figure 4b), which is similar to that generated for the

280 MTBC (Figure 3b) inasmuch as the mean R was continuously greater than one in the L4
 281 population since its emergence.
 282

Model	Mean Likelihood	Mean Rate (95% HPD)	Mean Rate Variance (95% HPD)	Mean Tree Height (95% HPD)	Origin (BDSKY only)
BDSKY+UCLD	-6032794.215	2.8369E-8 (1.4945E-8, 4.0535E-8)	4.071E-17 (5.4068E-18, 7.7999E-17)	1650.608 (945.7849, 2574.5831)	NA
BDSKY+UCLD+origin	-6032794.987	3.1477E-8 (2.0046E-8, 4.2189E-8)	4.5822E-17 (9.6022E-18, 8.051E-17)	1462.2611 (935.6968, 2058.31)	3268.341 (1102.2144, 8071.0277)
CC+UCLD	-6032797.480	3.1068E-8 (1.988E-8, 4.1624E-8)	4.3865E-17 (1.3291E-17, 7.806E-17)	1569.0512 (1054.607, 2225.4758)	NA
SKY+UCLD	-6032874.480	2.8097E-8 (1.5329E-8, 3.9927E-8)	3.7609E-17 (6.0593E-18, 7.1919E-17)	1690.536 (1016.2712, 2646.5163)	NA
CC+strict	-6033002.535	2.9299E-8 (2.2173E-8, 3.6637E-8)	NA	1567.544 (1186.1186, 1978.6488)	NA

283 **Table 3. Model comparison for L4 dataset.** Selected parameter estimates from five
 284 models applied to the Lineage 4 dataset: constant coalescent with uncorrelated lognormal
 285 clock (CC+UCLD), constant coalescent with strict clock (CC+strict), Bayesian skyline
 286 coalescent with uncorrelated lognormal clock (SKY+UCLD), birth-death skyline with
 287 uncorrelated lognormal clock and tree conditioned on the root (BDSKY+UCLD), and birth-
 288 death skyline with uncorrelated lognormal clock with origin parameter estimate
 289 (BDSKY+UCLD+origin).
 290



291

292 **Figure 4. L4 maximum clade credibility tree and birth-death skyline plot.** A) This MCC

293 tree of mean heights was generated from the BDSKY+UCLD model as applied to the L4

294 dataset. Modern genomes are collapsed according to sublineage (labeled on the right side).

295 The ancient genomes are labeled with their sample name. The Lineage 2 outgroup is

296 labeled as “L2_N0020.” The 95% HPD interval of selected node heights is indicated as

297 (lower boundary - upper boundary) in years before present. The time scale is expressed as

298 years before present, with the most recent time as 2010. B) The black line indicates median

299 reproductive number over time (reproductive number set to 5 dimensions, see Additional File

300 2). The shaded grey area represents the 95% HPD interval of the reproductive number. The

301 grey line indicates a reproductive number of 1. The red triangle on the timeline indicates the

302 position of LUND1.

303

304 DISCUSSION

305 The increasing number of ancient *Mycobacterium tuberculosis* genomes is steadily
306 reducing the uncertainty of molecular dating estimates for the emergence of the MTBC.
307 Here, using the ancient data available to date, we directly calibrate the MTBC time tree, and
308 confirm that known diversity within the complex is derived from a common ancestor that
309 existed ~2000-6000 years before present (Figure 3; Table 2) (5,6). Our results support the
310 hypothesis that the MTBC emerged during the Neolithic, and not before. The Neolithic
311 revolution generally refers to the worldwide transition in lifestyle and subsistence from more
312 mobile, foraging economies to more sedentary, agricultural economies made possible by the
313 domestication of plants and animals. The period during which it occurred varies between
314 regions. In Africa, where the MTBC is thought to have originated (3,39–41), the spread of
315 animal domestication in the form of pastoralism appears to have its focus around ~3000
316 BCE, or 5000 BP, across multiple regions (42). The estimates presented here place the
317 emergence of tuberculosis amidst the suite of human health impacts that took place as a
318 consequence of the Neolithic lifestyle changes often referred to collectively as the first
319 epidemiological transition (43,44).

320 Tuberculosis has left testaments to its history as a human pathogen in the
321 archaeological record (45), and some skeletal evidence has implied the existence of
322 tuberculosis in humans and animals pre-dating the lower 95% HPD boundary for the MTBC
323 MRCA presented here (7,8,10,46–50). However, it is important to explore the evolutionary
324 history of the MTBC through molecular data. Furthermore, it is crucial to base molecular
325 dating estimates on datasets that include ancient genomes, which expand the temporal
326 sampling window and provide data from the pre-antibiotic era. Numerous studies have found
327 long-term nucleotide substitution rate estimates in eukaryotes and viruses to be dependent
328 on the temporal breadth of the sampling window, and it is reasonable to assume the same
329 principle applies to bacteria (51–56). Additionally, rate variation over time and between
330 lineages, which may arise due to changing evolutionary dynamics such as climate and host

331 biology, can impact the constancy of the molecular clock (54,55). Though models have been
332 developed to accommodate uncertainty regarding these dynamics (57), temporally
333 structured populations can provide evidence and context for these phenomena over time
334 and can aid researchers in refining models appropriate for the taxon in question (56).

335 In addition to our MRCA estimate for the MTBC, we present one for L4, which is
336 among the most globally dominant lineages in the complex (31,58). Our analyses yielded
337 MRCA dates between ~1000-2500 years before present, as extrapolated from the 95% HPD
338 intervals of all models (Table 3), with the mean dates spanning from 320-548 CE. These
339 results are strikingly similar to those found in two prior publications, and support the idea
340 proposed by Kay and colleagues that L4 may have emerged during the late Roman period
341 (5,6). However, there exist discrepancies between different estimates for the age of this
342 lineage in available literature that touch the upper (37) and lower (58) edges of the 95%
343 HPD intervals reported here. In addition, recent phylogeographic analyses of the MTBC and
344 its lineages had ambiguous results for L4, with the internal nodes being assigned to either
345 African or European origins depending on the study or different dataset structures used
346 within the same study (37,58). Despite the ambiguity, this finding belies a close relationship
347 between ancestral L4 strains in Europe and Africa (37,58). Stucki and colleagues delineated
348 L4 into globally distributed “generalist” sublineages and highly local “specialist” sublineages
349 that do not appear outside a restricted geographical niche (31). Thus far, the specialist
350 sublineages are limited to the African continent; however, a clear phylogenetic relationship
351 explaining the distinction between geographically expansive and limited strains has not been
352 established. Specifically, LUND1 falls within the globally distributed, “generalist” L4.10/PGG3
353 sublineage that shares a clade with two specialist sublineages: L4.6.1/Uganda and
354 L4.6.2/Cameroon (Figure 4) (31). In the BDSKY+UCLD model presented here, the ancestral
355 node for this clade dates to approximately 1372 years BP. At this extrapolated time, an
356 ancestral strain underwent an evolutionary event in which some descendant lineages
357 acquired or lost a feature that equipped them to expand past limited host niches into
358 Eurasia. Confirming and elucidating this phenomenon could offer relevant clues regarding

359 the evolutionary relationship between populations of MTBC organisms and humans.
360 However, the current discrepancies over the age and geographic origin of L4 make
361 interpretations of existing data unreliable for this purpose. These discrepancies could be due
362 to differences in genome selection, SNP selection, and/or model selection and
363 parameterization. Until more diverse, high-quality ancient L4 genomes are generated,
364 creating a more temporally and geographically structured dataset, it is unlikely we will gain
365 clarity.

366 Going deeper into comparisons between the results presented here and those from
367 prior studies, mutation rate estimates in the L4 and full MTBC analyses were lower than
368 previous estimates for comparable datasets, but within the same order of magnitude, with all
369 mean and median estimates ranging between $1E-8$ and $5E-8$ (5,6) (Table 2). Nucleotide
370 substitution rates inferred based on modern tuberculosis data are close to, but slightly higher
371 than those based on ancient calibration, with multiple studies finding rates of approximately
372 $1E-7$ substitutions per site per year in multiple studies (4,59). Despite a strict clock model
373 having been rejected by the MEGA-CC molecular clock test (60) for both the L4 and full
374 MTBC datasets, the clock rate variation estimates do not surpass $9E-17$ in any model.
375 Additionally, there is little difference between the clock rates estimated in the L4 and full
376 MTBC datasets suggesting the rate of evolution in L4 does not meaningfully differ from that
377 of the full complex (Tables 2 and 3; Figure 5; Figures S9 and S10 in Additional File 2).

378 Another parameter explored here is R over time for the MTBC and L4 (Figures 2b
379 and 3b). For both datasets, we see an increase in R at approximately 750 BP. In the MTBC
380 model, it increases sharply and maintains its peak between 4 and 5. The increase is more
381 gradual in the L4 model, and declines to hovering just above $R=1$. This roughly coincides
382 with a jump in effective population size estimated by Liu and colleagues for MTBC lineages
383 indigenous to China (61). The decline of R for L4 beginning approximately 350 BP appears
384 surprising, given the historically recorded rise of the White Plague in Europe from the 17th-
385 19th century (62). However, this is likely due to the reduced sampling of modern sequences,

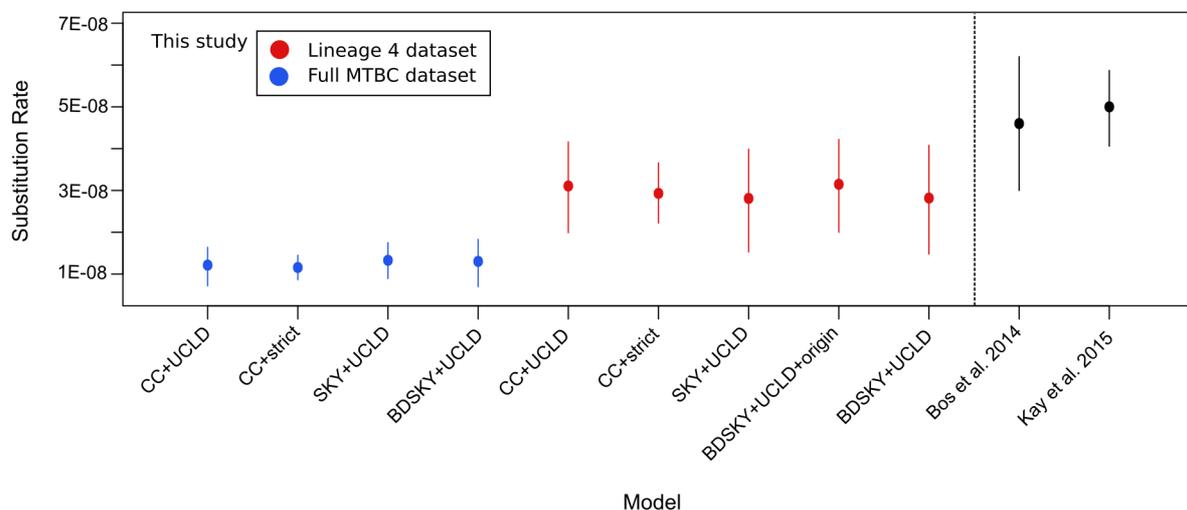
386 which enabled the dating of the entire L4 lineage. A thorough phylodynamic analysis
387 requires inclusion of “outbreak samples” only (63) and shall be explored in future work.

388 Importantly, we explored our data through multiple models, including birth-death tree
389 priors. In our opinion, these models offer more robust parameterization options for
390 heterochronous datasets that are unevenly distributed over time, such as those presented
391 here, by allowing for uneven sampling proportions across different time intervals of the tree
392 (64). Recent studies have demonstrated the importance of selecting appropriate tree priors
393 for the population under investigation, as well as the differences between birth-death and
394 coalescent tree priors (65,66). It is notable that the estimates reported here roughly agree
395 across multiple demographic and clock models implemented in BEAST2. The estimate of
396 the origin height for the L4 dataset as calculated with the birth-death Skyline model overlaps
397 with the 95% HPD intervals for the tree height estimates across models in the full MTBC
398 dataset.

399 In addition to confirming the findings of prior publications, this study contributes a
400 high-coverage, contamination-free, and securely dated ancient *M. tuberculosis* genome for
401 future dating efforts, which may include more ancient data or more realistic models. Much of
402 this quality likely comes from the unique preservation environment of the calcified nodule. In
403 the case of tuberculosis, such nodules form from host immunological responses in the
404 waning period of an active pulmonary infection and remain in lung tissue, characterizing the
405 latent form of the disease. Host immune cells were likely responsible for the dominant signal
406 of 89% human DNA in the LUND1 metagenomic screening library. Similar levels of
407 preservation have been observed through analyses of ancient nodules yielding *Brucella*
408 (Kay et al, 2014) and urogenital bacterial infections (DeVault et al, 2017), with pathogen
409 preservation rivaling what we report here.

410 LUND1 avoided multiple quality-related problems often encountered in the
411 identification and reconstruction of ancient genetic data from the MTBC. The genome is of
412 high quality both in terms of its high coverage and low heterozygosity. Despite the low
413 quantity of MTBC DNA detected in the preliminary screening data, in-solution capture

414 enriched the proportion of endogenous DNA by three orders of magnitude (Table 1). The
415 resultant genomic coverage left few ambiguous positions at which multiple alleles were
416 represented by greater than 10% of the aligned reads. This extremely low level of
417 heterozygosity indicated that LUND1 contained a dominant signal of only one MTBC strain.
418 This circumvented analytical complications that can arise from the simultaneous presence of
419 multiple MTBC strains associated with mixed infections, or from the presence of abundant
420 non-MTBC mycobacteria stemming from the environment. The preservation conditions of
421 Bishop Winstrup's remains, mummified in a crypt far from soil, left the small MTBC signal
422 unobscured by environmental mycobacteria or by the dominance of any other bacterial
423 organisms (Figure 2a). The unprecedented quality of LUND1 and the precision of its
424 calibration point (historically recorded year of death) made it ideal for Bayesian molecular
425 dating applications.
426



427
428 **Figure 5. Substitution rate comparison across models and studies.** Mean substitution
429 rate per site per year for all models is expressed by a filled circle, with extended lines
430 indicating the 95% HPD interval for that parameter. The Bos et al. 2014 and Kay et al. 2015
431 ranges are based on the reported rate values in each study. The Bos et al. 2014 range is
432 based on a full MTBC dataset, while the Kay et al. 2015 range is based on an L4 dataset. All
433 values presented here fall within one order of magnitude.

434

435 As the practice of applying ancient genomic data is still in its nascent stages, there
436 are caveats to the results of this study. First, this analysis excludes *M. canettii* – a bacterium
437 that can cause pulmonary tuberculosis – from the MTBC dataset, and as such our estimate
438 does not preclude the possibility of a closely related ancestor having caused tuberculosis-
439 like infections in humans before 6,000 BP. The inferred MRCA could be restricted to a
440 lineage that survived an evolutionary bottleneck, possibly connected to its virulence in
441 humans as suggested elsewhere, albeit as a considerably more ancient event (67,68).
442 Additionally, despite the use of ancient data, our temporal sampling window is still narrow
443 given the estimated age of the MTBC and L4. For the MTBC dataset no samples pre-date
444 1,000 years before present, and for L4, no samples predate 350 years before present. It
445 could be argued the ancient L4 genomes available to date represent samples taken in the
446 midst of an epidemic – namely, the “White Plague” of tuberculosis, which afflicted Europe
447 between the 17th and 19th centuries (62). For a slow-evolving bacterial pathogen like
448 tuberculosis, it is possible our sampling window of ancient genomes is subject to the very
449 issue they are meant to alleviate: the time-dependency of molecular clocks (51,53–55). The
450 genomes sampled from pre-contact Peruvian remains do not derive from a known epidemic
451 period in history and add temporal spread to our MTBC dataset. However, their membership
452 to a clade of animal-associated strains (*M. pinnipedii*) indicates they were subject to
453 dramatically different evolutionary pressures compared to the human-associated lineages of
454 the complex due to differing host biology and population dynamics. On a related matter, the
455 available ancient MTBC genomes also suffer from a lack of lineage diversity, with only
456 pinniped strains and L4 represented.

457 Filling the MTBC time tree with more ancient genomes from diverse time periods,
458 locations, and lineages would address the limitations listed above. The most informative
459 data would a) derive from an Old World context (i.e. Europe, Asia, or Africa) pre-dating the
460 White Plague in Europe or b) come from any geographical location or pre-modern time
461 period, but belong to one of the MTBC lineages not yet represented by ancient data. An

462 ideal data point, which would clarify many open questions and seeming contradictions
463 related to the evolutionary history of the MTBC, would derive from Africa, the inferred home
464 of the MTBC ancestor (3,39–41), and pre-date 2,000 years before present. A genome of this
465 age would test the lower boundaries of the 95% HPD tree height intervals estimated in the
466 full MTBC models presented here. Until recently it would have been considered unrealistic to
467 expect such data to be generated from that time period and location. Innovations and
468 improvements in ancient DNA retrieval and enrichment methods, however, have brought this
469 expectation firmly into the realm of the possible (30,69). Ancient bacterial pathogen
470 genomes have now been retrieved from remains from up to 5,000 years before present (70–
471 72) and recent studies have reported the recovery of human genomes from up to 15,000
472 year-old remains from north Africa (73,74).

473 CONCLUSIONS

474 Here we offer confirmation that the extant MTBC, and all available ancient MTBC
475 genomes, stem from a common ancestor that existed a maximum of 6,000 years before
476 present. Many open questions remain, however, regarding the evolutionary history of the
477 MTBC and its constituent lineages, as well as the role of tuberculosis in human history.
478 Elucidating these questions is an iterative process, and progress will include the generation
479 of diverse ancient *M. tuberculosis* genomes, and the refinement and improved
480 parameterization of Bayesian models that reflect the realities of MTBC (and other
481 organisms') population dynamics and sampling frequencies over time. To aid in future
482 attempts to answer these questions, this study provides an ancient MTBC genome of
483 impeccable quality and explores the first steps in applying birth-death population models to
484 modern and ancient TB data.

485 METHODS

486 **Lung nodule identification**

487 The paleopathological investigation of the body of Winstrup is based on extensive
488 CT-scan examinations with imaging of the mummy and its bedding performed with a
489 Siemens Somatom Definition Flash, 128 slice at the Imaging Department of Lund University

490 Hospital. Ocular inspection of the body other than of the head and hands was not feasible,
491 since Winstrup was buried in his episcopal robes and underneath the body was wrapped in
492 linen strips. The velvet cap and the leather gloves were removed during the
493 investigation. The body was naturally mummified and appeared to be well preserved with
494 several internal organs identified.

495 The imaging was quite revealing. The intracranial content was lost with remains of
496 the brain in the posterior skull base. Further, the dental status was poor with several teeth in
497 the upper jaw affected by severe attrition, caries and signs of tooth decay, as well as the
498 absence of all teeth in the lower jaw. Most of the shed teeth were represented by closed
499 alveoli, indicating antemortem tooth loss. Along with the investigation of the bedding, a small
500 sack made of fabric was found behind the right elbow containing five teeth: two incisors, two
501 premolars and one molar. The teeth in the bag complemented the remaining teeth in the
502 upper jaw. It is feasible that the teeth belonged to Winstrup and were shed several years
503 before he died. A fetus approximately five months of age was also found in the bedding,
504 underneath his feet.

505 Both lungs were preserved but collapsed with findings of a small parenchymal
506 calcification and two ~5 mm calcifications in the right hilum (Figure 1). The assessment was
507 that these could constitute a Ranke complex, suggestive of previous primary tuberculosis. A
508 laparoscopy was performed at the Lund University Hospital in a clinical environment
509 whereby the nodules were retrieved. Furthermore, several calcifications were also found in
510 the aorta and the coronary arteries, suggesting the presence of atherosclerosis. The
511 stomach, liver and gall bladder were preserved, and several small gallstones were observed.
512 The spleen could be identified but not the kidneys. The intestines were there, however,
513 collapsed except for the rectum that contained several large pieces of concretions. The
514 bladder and the prostate could not be recognized.

515 The skeleton showed several pathological changes. Findings on the vertebrae
516 consistent with of DISH (Diffuse idiopathic skeletal hyperostosis) were present in the
517 thoracic and the lumbar spine. Reduction of the joint space in both hip joints and the left

518 knee joint indicate that Winstrup was affected by osteoarthritis. No signs of gout or
519 osteological tuberculosis (i.e. Pott's disease) were found.

520 Neither written sources nor the modern examination of the body of Winstrup reveal
521 the immediate cause of death. However, it is known that he was bedridden for at least two
522 years preceding his death. Historical records indicate that gallstones caused him problems
523 while travelling to his different parishes. Additionally, he was known to have suffered from
524 tuberculosis as a child, which may have recurred in his old age.

525 **Sampling and extraction**

526 Sampling of the lung nodule, extraction, and library preparation were conducted in
527 dedicated ancient DNA clean rooms at the Max Planck Institute for the Science of Human
528 History in Jena, Germany. The nodule was broken using a hammer, and a 5.5 mg portion of
529 the nodule was taken with lung tissue for extraction according to a previously described
530 protocol with modifications (18). The sample was first decalcified overnight at room
531 temperature in 1 mL of 0.5 M EDTA. The sample was then spun down, and the EDTA
532 supernatant was removed and frozen. The partially decalcified nodule was then immersed in
533 1 mL of a digestion buffer with final concentrations of 0.45 M EDTA and 0.25 mg/mL
534 Proteinase K (Qiagen) and rotated at 37°C overnight. After incubation, the sample was
535 centrifuged. The supernatants from the digestion and initial decalcification step were purified
536 using a 5 M guanidine-hydrochloride binding buffer with a High Pure Viral Nucleic Acid Large
537 Volume kit (Roche). The extract was eluted in 100 µl of a 10mM tris-hydrochloride, 1 mM
538 EDTA (pH 8.0), and 0.05% Tween-20 buffer (TET). Two negative controls and one positive
539 control sample of cave bear bone powder were processed alongside LUND1 to control for
540 reagent/laboratory contamination and process efficiency, respectively.

541 **Library preparation and shotgun screening sequencing**

542 Double-stranded Illumina libraries were constructed according to an established
543 protocol with some modifications (75). Overhangs of DNA fragments were blunt-end
544 repaired in a 50 µl reaction including 10 µl of the LUND1 extract, 21.6 µl of H₂O, 5 µl of NEB
545 Buffer 2 (New England Biolabs), 2 µl dNTP mix (2.5 mM), 4 µl BSA (10 mg/ml), 5 µl ATP (10

546 mM), 2 μ l T4 polynucleotide kinase, and 0.4 μ l T4 polymerase, then purified and eluted in 18
547 μ l TET. Illumina adapters were ligated to the blunt-end fragments in a reaction with 20 μ l
548 Quick Ligase Buffer, 1 μ l of adapter mix (0.25 μ M), and 1 μ l of Quick Ligase. Purification of
549 the blunt-end repair and adapter ligation steps was performed using MinElute columns
550 (Qiagen). Adapter fill-in was performed in a 40 μ l reaction including 20 μ l adapter ligation
551 eluate, 12 μ l H₂O, 4 μ l Thermopol buffer, 2 μ l dNTP mix (2.5 mM), and 2 μ l Bst polymerase.
552 After the reaction was incubated at 37°C for 20 minutes, the enzyme was heat deactivated
553 with a 20 minute incubation at 80°C. Four library blanks were processed alongside LUND1
554 to control for reagent/laboratory contamination. The library was quantified using a real-time
555 qPCR assay (Lightcycler 480 Roche) with the universal Illumina adapter sequences IS7 and
556 IS8 as targets. Following this step, the library was double indexed (76) with a unique pair of
557 indices over two 100 μ l reactions using 19 μ l of template, 63.5 μ l of H₂O, 10 μ l PfuTurbo
558 buffer, 1 μ l PfuTurbo (Agilent), 1 μ l dNTP mix (25mM), 1.5 μ l BSA (10 mg/ml), and 2 μ l of
559 each indexing primer (10 μ M). The master mix was prepared in a pre-PCR clean room and
560 transported to a separate lab for amplification. The two reactions were purified and eluted in
561 25 μ l of TET each over MinElute columns (Qiagen), then assessed for efficiency using a
562 real-time qPCR assay targeting the IS5 and IS6 sequences in the indexing primers. The
563 reactions were then pooled into one double-indexed library. Approximately one-third of the
564 library was amplified over three 70 μ l PCR reactions using 5 μ l of template each and
565 Herculase II Fusion DNA Polymerase (Agilent). The products were MinElute purified, pooled,
566 and quantified using an Agilent Tape Station D1000 Screen Tape kit. LUND1 and the
567 corresponding negative controls were sequenced separately on an Illumina NextSeq 500
568 using single-end, 75-cycle, high-output kits.

569 **Pathogen identification and authentication**

570 De-multiplexed sequencing reads belonging to LUND1 were processed *in silico* with
571 the EAGER pipeline (v.1.92) (24). ClipAndMerge was used for adapter removal, fragment
572 length filtering (minimum sequence length: 30 bp), and base sequence quality filtering
573 (minimum base quality: 20). MALT v. 038 (19) was used to screen the metagenomic data for

574 pathogens using the full NCBI Nucleotide database ('nt', April 2016) with a minimum percent
575 identity of 85%, a minSupport threshold of 0.01, and a topPercent value of 1.0. The resulting
576 metagenomic profile was visually assessed with MEGAN6 CE (20). The adapter-clipped
577 reads were additionally aligned to a reconstructed MTBC ancestor genome (23) with BWA
578 (33) as implemented in EAGER (-l 1000, -n 0.01, -q 30). Damage was characterized with
579 DamageProfiler in EAGER (77).

580 **In-solution capture probe design**

581 Single-stranded probes for in-solution capture were designed using a computationally
582 extrapolated ancestral genome of the MTBC (23). The probes are 52 nucleotides in length
583 with a tiling density of 5 nucleotides, yielding a set of 852,164 unique probes after the
584 removal of duplicate and low complexity probes. The number of probes was raised to
585 980,000 by a random sampling among the generated probe sequences. A linker sequence
586 (5'-CACTGCGG-3') was attached to each probe sequence, resulting in probes of 60
587 nucleotides in length, which were printed on a custom-design 1 million-feature array
588 (Agilent). The printed probes were cleaved off the array, biotinylated and prepared for
589 capture according to Fu et al. (30).

590 **UDG library preparation and in-solution capture**

591 Fifty microliters of the original LUND1 extract were used to create a uracil-DNA
592 glycosylase (UDG) treated library, in which the post-mortem cytosine to uracil modifications,
593 which cause characteristic damage patterns in ancient DNA, are removed. The template
594 DNA was treated in a buffer including 7 μ l H₂O, 10 μ l NEB Buffer 2 (New England Biolabs),
595 12 μ l dNTP mix (2.5 mM), 1 μ l BSA (10 mg/ml), 10 μ l ATP (10 mM), 4 μ l T4 polynucleotide
596 kinase, and 6 μ l USER enzyme (New England Biolabs). The reaction was incubated at 37°C
597 for three hours, then 4 μ l of T4 polymerase was added to the library to complete the blunt-
598 end repair step. The remainder of the library preparation protocol, including double indexing,
599 was performed as described above.

600 The LUND1 UDG-treated library was amplified over two rounds of amplification using
601 Herculase II Fusion DNA Polymerase (Agilent). In the first round, five reactions using 3 μ l of

602 template each were MinElute purified and pooled together. The second round of
603 amplification consisted of three reactions using 3 μ l of template each from the first
604 amplification pool. The resulting products were MinElute purified and pooled together. The
605 final concentration of 279 ng/ μ l was measured using an Agilent Tape Station D1000 Screen
606 Tape kit (Agilent). A portion of the non-UDG library (see above) was re-amplified to 215
607 ng/ μ l. A 1:10 pool of the non-UDG and UDG amplification products was made to undergo
608 capture. A pool of all associated negative control libraries (Supplementary Table 2) and a
609 positive control known to contain *M. tuberculosis* DNA also underwent capture in parallel
610 with the LUND1 libraries. Capture was performed according to an established protocol (29),
611 and the sample product was sequenced on an Illumina HiSeq 4000 with a 150-cycle paired
612 end kit to a depth of ~60 million paired reads. The blanks were sequenced on a NextSeq
613 500 with a 75-cycle paired end kit.

614 **Genomic reconstruction, heterozygosity, and SNP calling**

615 For the enriched, UDG-treated LUND1 sequencing data, de-multiplexed paired-end
616 reads were processed with the EAGER pipeline (v. 1.92) (24), adapter-clipped with
617 AdapterRemoval, and aligned to the MTBC reconstructed ancestor genome with in-pipeline
618 BWA (-l 32, -n 0.1, -q 37). Previously published ancient and modern *Mycobacterium*
619 *tuberculosis* genomic data (Supplementary Table 4, Supplementary Table 5) were
620 processed as single-end sequencing reads, but otherwise processed identically in the
621 EAGER pipeline. Genome Analysis Toolkit (GATK) UnifiedGenotyper was used to call SNPs
622 using default parameters and the EMIT ALL SITES output option (78). We used
623 MultiVCFAnalyzer (v0.87 <https://github.com/alexherbig/MultiVCFAnalyzer>) (5) to create and
624 curate SNP alignments for the L4 (Supplementary Table 5) and full MTBC (Supplementary
625 Table 4) datasets based on SNPs called in reference to the TB ancestor genome (23), with
626 repetitive sequences, regions subject to cross-species mapping, and potentially imported
627 sites excluded. The repetitive and possibly cross-mapped regions were excluded as
628 described previously (5). Potentially imported sites were identified using ClonalFrameML
629 (34) separately for each dataset, using full genomic alignments and trees generated in

630 RAxML (79) as input without the respective outgroups. Remaining variants were called as
631 homozygous if they were covered by at least 5 reads, had a minimum genotyping quality of
632 30, and constituted at least 90% of the alleles present at the site. Outgroups for each
633 dataset were included in the SNP alignments, but no variants unique to the selected
634 outgroup genomes were included. Minority alleles constituting over 10% were called and
635 assessed for LUND1 to check for a multiple strain *M. tuberculosis* infection. Sites with
636 missing or incomplete data were excluded from further analysis.

637 **Phylogenetic analysis**

638 Maximum likelihood, maximum parsimony, and neighbor joining trees were
639 generated for the L4 and full MTBC datasets (Tables S4 and S5 in Additional File 1), with
640 500 bootstrap replications per tree. Maximum parsimony and neighbor joining trees were
641 configured using MEGA-Proto and executed using MEGA-CC (60). Maximum likelihood
642 trees were configured and executed using RAxML (79) with the GTR+GAMMA substitution
643 model.

644 **Bayesian phylogenetic analysis of full MTBC and L4 datasets**

645 Bayesian phylogenetic analysis of the full MTBC was conducted using a dataset of
646 261 *M. tuberculosis* genomes including LUND1, five previously published ancient genomes
647 (5,6), and 255 previously published modern genomes (Table S4 in Additional File 1).
648 *Mycobacterium canettii* was used as an outgroup for this dataset. Bayesian phylogenetic
649 analysis of L4 of the MTBC was conducted using a dataset of 152 genomes including three
650 ancient genomes presented here and in a previous publication (6) and 149 previously
651 published modern genomes (Table S5 in Additional File 1). Body80 and body92 were
652 selected out of the eight samples presented by Kay and colleagues based on multiple
653 criteria. Multiple samples from that study proved to be mixed strain infections. Apart from
654 body92, these samples were excluded from this analysis due to our present inability to
655 separate strains without ignoring derived positions. Body92 had a clearly dominant strain
656 estimated by Kay et al. (6) to make up 96% of the tuberculosis data, and stringent mapping
657 in BWA (33) (-l 32, -n 0.1, -q 37) found the genome to have 124-fold coverage when mapped

658 against the TB ancestor. Between the degree of dominance and the high coverage, we
659 could confidently call variant positions from the dominant strain (Figure S8a in Additional File
660 2). Body80 was the only single-strain sample from that collection to have sufficient coverage
661 (~8x) for confident SNP calling after stringent mapping (Figure S8b in Additional File 2). For
662 selection criteria for the modern genomes, please see Additional File 2. L2_N0020 was used
663 as an outgroup. The possibility of equal evolutionary rates in both datasets was rejected by
664 the MEGA-CC molecular clock test (60). TempEst (80) was also used to assess temporal
665 structure in the phylogeny prior to analysis with BEAST2 (36) (full MTBC $R^2=0.273$; L4
666 $R^2=0.113$).

667 A correction for static positions in the *M. tuberculosis* genome not included in the
668 SNP alignment was included in the configuration file. A “TVM” substitution model, selected
669 based on results from ModelGenerator (81), was implemented in BEAUti as a GTR+G4
670 model with the AG rate parameter fixed to 1.0. LUND1, body80, and body92 were tip-
671 calibrated using year of death, which was available for all three individuals (Table S5 in
672 Additional File 1). The three ancient Peruvian genomes were calibrated using the mid-point
673 of their OxCal ranges (Table S4 in Additional File 1) (5). We performed tip sampling for all
674 modern genomes excluding the outgroup over a uniform distribution between 1992 and 2010
675 for all but the BDSKY models for both datasets. The outgroup was fixed to 2010 in every
676 case. In the BDSKY models, all modern genomes were given a tip date of 2010. All tree
677 priors were used in conjunction with an uncorrelated relaxed lognormal clock model. The
678 constant coalescent model was also used in conjunction with a strict clock model.

679 Two independent MCMC chains of 200,000,000 iterations minimum were computed
680 for each model. If the ESS for any parameter was below 200 after the chains were
681 combined, they were resumed with additional iterations. The results were assessed in
682 Tracer v1.7.1 with a 10 percent burn-in (82). Trees were sampled every 20,000 iterations.
683 The log files and trees for each pair of runs were combined using LogCombiner v2.4.7 (36).
684 An MCC tree was generated using TreeAnnotator with ten percent burn-in (36). For details
685 on the parameterization of the birth-death models, please see Additional File 2.

686

687 **DECLARATIONS**

688 **Ethics approval and consent to participate**

689 Not applicable

690 **Consent for publication**

691 Not applicable

692 **Availability of data and material**

693 Raw sequencing data from the non-UDG, non-enriched screening library and the UDG-
694 treated, enriched library can be found under the BioProject PRJNA517266.

695 **Competing interests**

696 Not applicable

697 **Funding**

698 This project was supported by the Max Planck Society and by grants from the Erik Philip-
699 Sørensen Foundation and Crafoord foundation to T.A. and C.A.

700 **Authors' contributions**

701 C.A. and K.I.B. conceived of the investigation. S.S., D.K., A.H., and K.I.B. designed the
702 experiments. T.A., G.B., and C.A. performed the exhumation and radiological analysis of the
703 mummy and provided a paleopathological examination. G.B. was responsible for the CT
704 examinations together with imaging analysis and coordination of the calcification extraction.
705 S.S. performed laboratory work. S.S., D.K., A.H., Å.J.V., and K.I.B. performed analyses.

706 **Acknowledgments**

707 The authors would like to thank Marta Burri for conducting the hybridization capture. We also
708 thank Elizabeth A. Nelson for laboratory assistance, Maria Spyrou and Felix M. Key
709 technical assistance, and all members of the Molecular Palaeopathology and Computational
710 Pathogenomics research groups at the MPI-SHH for constructive comments throughout the
711 study. We would like to acknowledge the Department of Imaging and Clinical Physiology,
712 Skåne University Hospital Lund for the generous opportunity to examine the mummy of
713 Peder Winstrup and the radiologists Roger Siemund, Pär Wingren, Mats Geijer, Pernilla

714 Gustavson and David Pellby for contributing to the image analyses. A sincere thank you to
715 the thoracic surgeons Erik Gyllstedt and Jesper Andreasson for so delicately extracting the
716 small calcification of interest and enabling further investigations.
717

718 REFERENCES

- 719 1. WHO. WHO | Tuberculosis (TB) [Internet]. WHO. 2018 [cited 2018 Nov 18]. Available
720 from: <http://www.who.int/tb/en/>
- 721 2. Houben RMGJ, Dodd PJ. The Global Burden of Latent Tuberculosis Infection: A Re-
722 estimation Using Mathematical Modelling. *PLOS Medicine*. 2016 Oct
723 25;13(10):e1002152.
- 724 3. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa
725 migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern
726 humans. *Nat Genet*. 2013 Oct;45(10):1176–82.
- 727 4. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, et al. The
728 Role of Selection in Shaping Diversity of Natural *M. tuberculosis* Populations. *PLOS*
729 *Pathogens*. 2013 Aug 15;9(8):e1003543.
- 730 5. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian
731 mycobacterial genomes reveal seals as a source of New World human tuberculosis.
732 *Nature*. 2014 Aug 20;514(7523):494–7.
- 733 6. Kay GL, Sergeant MJ, Zhou Z, Chan JZ-M, Millard A, Quick J, et al. Eighteenth-century
734 genomes show that mixed infections were common at time of peak tuberculosis in
735 Europe. *Nature Communications*. 2015 Apr 7;6:6717.
- 736 7. Baker O, Lee OY-C, Wu HHT, Besra GS, Minnikin DE, Llewellyn G, et al. Human
737 tuberculosis predates domestication in ancient Syria. *Tuberculosis*. 2015 Jun 1;95:S4–
738 12.
- 739 8. Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OY-C, Gernaey AM, et al.
740 Detection and Molecular Characterization of 9000-Year-Old *Mycobacterium*
741 *tuberculosis* from a Neolithic Settlement in the Eastern Mediterranean. Ahmed N,
742 editor. *PLoS ONE*. 2008 Oct 15;3(10):e3426.
- 743 9. Masson M, Molnár E, Donoghue HD, Besra GS, Minnikin DE, Wu HHT, et al.
744 Osteological and Biomolecular Evidence of a 7000-Year-Old Case of Hypertrophic
745 Pulmonary Osteopathy Secondary to Tuberculosis from Neolithic Hungary. *PLOS ONE*.
746 2013;8(10):e78252.
- 747 10. Rothschild BM, Martin L, Lev G, Bercovier H, Bar-Gal GK, Greenblatt CL, et al.
748 *Mycobacterium tuberculosis* Complex DNA from an Extinct Bison Dated 17,000 Years
749 before the Present. *Clinical Infectious Diseases*. 2001;(33):305–11.
- 750 11. Wilbur AK, Bouwman AS, Stone AC, Roberts CA, Pfister L-A, Buikstra JE, et al.
751 Deficiencies and challenges in the study of ancient tuberculosis DNA. *Journal of*
752 *Archaeological Science*. 2009 Sep;36(9):1990–7.
- 753 12. Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews*
754 *Microbiology*. 2018 Apr;16(4):202–13.
- 755 13. Ochman H, Elwyn S, Moran NA. Calibrating bacterial evolution. *PNAS*. 1999 Oct
756 26;96(22):12638–43.

- 757 14. Warinner C, Herbig A, Mann A, Yates JAF, Weiß CL, Burbano HA, et al. A Robust
758 Framework for Microbial Archaeology. *Annual Review of Genomics and Human*
759 *Genetics*. 2017;18(1).
- 760 15. Leung AN, Müller NL, Pineda PR, FitzGerald JM. Primary tuberculosis in childhood:
761 radiographic manifestations. *Radiology*. 1992 Jan;182(1):87–91.
- 762 16. Burrill J, Williams CJ, Bain G, Conder G, Hine AL, Misra RR. Tuberculosis: a radiologic
763 review. *Radiographics*. 2007 Oct;27(5):1255–73.
- 764 17. Bahr NC, Antinori S, Wheat LJ, Sarosi GA. Histoplasmosis infections worldwide:
765 thinking outside of the Ohio River valley. *Curr Trop Med Rep*. 2015 Jun 1;2(2):70–80.
- 766 18. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete
767 mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from
768 ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*. 2013
769 Sep 24;110(39):15758–63.
- 770 19. Vågene ÅJ, Herbig A, Campana MG, García NMR, Warinner C, Sabin S, et al.
771 *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in
772 Mexico. *Nature Ecology & Evolution*. 2018 Jan 15;1.
- 773 20. Huson DH, Beier S, Flade I, Górská A, El-Hadidi M, Mitra S, et al. MEGAN Community
774 Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing
775 Data. *PLoS Comput Biol*. 2016 Jun;12(6):e1004957.
- 776 21. Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, Arlat M, et al. Genome
777 sequence of the plant pathogen *Ralstonia solanacearum*. *Nature*. 2002
778 Jan;415(6871):497–502.
- 779 22. Mann AE, Sabin S, Ziesemer K, Vågene ÅJ, Schroeder H, Ozga AT, et al. Differential
780 preservation of endogenous human and microbial DNA in dental calculus and dentin.
781 *Scientific Reports*. 2018 Jun 29;8(1):9822.
- 782 23. Comas I, Chakravarti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T
783 cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat*
784 *Genet*. 2010 Jun;42(6):498–503.
- 785 24. Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, et al. EAGER: efficient
786 ancient genome reconstruction. *Genome Biology*. 2016;17:60.
- 787 25. Renaud G, Slon V, Duggan AT, Kelso J. Schmutzi: estimation of contamination and
788 endogenous mitochondrial consensus calling for ancient DNA. *Genome Biology*.
789 2015;16(1):1–18.
- 790 26. Dabney J, Meyer M, Pääbo S. Ancient DNA Damage. *Cold Spring Harb Perspect Biol*.
791 2013 Jul 1;5(7):a012567.
- 792 27. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L. mapDamage: testing
793 for damage patterns in ancient DNA sequences. *Bioinformatics*. 2011 Aug
794 1;27(15):2153–5.
- 795 28. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of
796 deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic*
797 *Acids Res*. 2010 Apr 1;38(6):e87–e87.

- 798 29. Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, Brizuela L, et al.
799 Hybrid selection of discrete genomic intervals on custom-designed microarrays for
800 massively parallel sequencing. *Nat Protocols*. 2009 May;4(6):960–74.
- 801 30. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, et al. DNA analysis of an
802 early modern human from Tianyuan Cave, China. *PNAS*. 2013 Feb 5;110(6):2223–7.
- 803 31. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium
804 tuberculosis lineage 4 comprises globally distributed and geographically restricted
805 sublineages. *Nat Genet*. 2016 Oct 31;48:1535–43.
- 806 32. Coll F, McNERney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A
807 robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nature*
808 *Communications*. 2014 Sep 1;5:4812.
- 809 33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
810 transform. *Bioinformatics*. 2009 Jul 15;25(14):1754–60.
- 811 34. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole
812 Bacterial Genomes. *PLOS Computational Biology*. 2015 Feb 12;11(2):e1004041.
- 813 35. Boritsch EC, Khanna V, Pawlik A, Honoré N, Navas VH, Ma L, et al. Key experimental
814 evidence of chromosomal DNA transfer among selected tuberculosis-causing
815 mycobacteria. *PNAS*. 2016 Aug 15;201604921.
- 816 36. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: A
817 Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology*.
818 2014 Apr 10;10(4):e1003537.
- 819 37. O'Neill MB, Shockey AC, Zarley A, Aylward W, Eldholm V, Kitchen A, et al. Lineage
820 specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia.
821 *bioRxiv*. 2018 Jul 6;210161.
- 822 38. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for
823 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*.
824 2012 Apr 1;6(2):80–92.
- 825 39. Wirth T, Hildebrand F, Allix-Béguec C, Wölbelling F, Kubica T, Kremer K, et al. Origin,
826 Spread and Demography of the Mycobacterium tuberculosis Complex. *PLOS*
827 *Pathogens*. 2008 Sep 26;4(9):e1000160.
- 828 40. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High
829 Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human
830 Demography. *PLOS Biology*. 2008 Dec 16;6(12):e311.
- 831 41. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omais B, Marmiesse M, et al. Ancient
832 Origin and Gene Mosaicism of the Progenitor of Mycobacterium tuberculosis. *PLOS*
833 *Pathogens*. 2005 Aug 19;1(1):e5.
- 834 42. Shillington K. *History of Africa*. Third. New York: Palgrave MacMillan; 2012.
- 835 43. Cohen MN, Armelagos GJ. *Paleopathology at the Origins of Agriculture*. Gainesville,
836 Florida: University Press of Florida; 1984.

- 837 44. Armelagos GJ, Brown PJ, Turner B. Evolutionary, historical and political economic
838 perspectives on health and disease. *Social Science & Medicine*. 2005 Aug;61(4):755–
839 65.
- 840 45. Roberts CA, Buikstra JE. *The Bioarchaeology of Tuberculosis: A Global View on a*
841 *Reemerging Disease*. Gainesville, Florida: University Press of Florida; 2003.
- 842 46. Canci A, Minozzi S, Tarli SMB. New Evidence of Tuberculous Spondylitis from Neolithic
843 Liguria (Italy). *International Journal of Osteoarchaeology*. 1996 Dec 1;6(5):497–501.
- 844 47. El-Najjar M, Al-Shiyab A, Al-Sarie I. Cases of tuberculosis at 'Ain Ghazal, Jordan.
845 *Paléorient*. 1996;22(2):123–8.
- 846 48. Formicola V, Milanese Q, Scarsini C. Evidence of spinal tuberculosis at the beginning of
847 the fourth millennium BC from Arene Candide cave (Liguria, Italy). *American Journal of*
848 *Physical Anthropology*. 1987 Jan 1;72(1):1–6.
- 849 49. Köhler K, Pálfi G, Molnár E, Zalai-Gaál I, Oszás A, Bánffy E, et al. A Late Neolithic
850 Case of Pott's Disease from Hungary. *International Journal of Osteoarchaeology*. 2014
851 Nov 1;24(6):697–703.
- 852 50. Sparacello VS, Roberts CA, Kerudin A, Müller R. A 6500-year-old Middle Neolithic child
853 from Pollera Cave (Liguria, Italy) with probable multifocal osteoarticular tuberculosis.
854 *International Journal of Paleopathology* [Internet]. 2017 Feb [cited 2017 Feb 13];
855 Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1879981716300900>
- 856 51. Ho SYW, Phillips MJ, Cooper A, Drummond AJ. Time Dependency of Molecular Rate
857 Estimates and Systematic Overestimation of Recent Divergence Times. *Mol Biol Evol*.
858 2005 Jul 1;22(7):1561–8.
- 859 52. Ho SYW, Larson G. Molecular clocks: when times are a-changin'. *Trends in Genetics*.
860 2006 Feb;22(2):79–83.
- 861 53. Ho SYW, Shapiro B, Phillips MJ, Cooper A, Drummond AJ. Evidence for Time
862 Dependency of Molecular Rate Estimates. *Syst Biol*. 2007 Jun 1;56(3):515–22.
- 863 54. Achtman M. How old are bacterial pathogens? *Proc R Soc B*. 2016 Aug
864 17;283(1836):20160990.
- 865 55. Duchêne S, Holmes EC, Ho SYW. Analyses of evolutionary dynamics in viruses are
866 hindered by a time-dependent bias in rate estimates. *Proc R Soc B*. 2014 Jul
867 7;281(1786):20140732.
- 868 56. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving
869 populations. *Trends in Ecology & Evolution*. 2003 Sep;18(9):481–8.
- 870 57. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating
871 with Confidence. *PLOS Biology*. 2006 Mar 14;4(5):e88.
- 872 58. Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, et al.
873 Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial
874 migration and local adaptation. *Science Advances*. 2018 Oct 1;4(10):eaat5869.

- 875 59. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, et al. Use of whole
876 genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis*
877 during latent infection. *Nature Genetics*. 2011 May;43(5):482–6.
- 878 60. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: computing core of molecular
879 evolutionary genetics analysis program for automated and iterative data analysis.
880 *Bioinformatics*. 2012 Oct 15;28(20):2685–6.
- 881 61. Liu Q, Ma A, Wei L, Pang Y, Wu B, Luo T, et al. China's tuberculosis epidemic stems
882 from historical expansion of four strains of *Mycobacterium tuberculosis*. *Nature Ecology*
883 & *Evolution*. 2018 Nov 5;1.
- 884 62. Dubos R, Dubos J. *The White Plague: Tuberculosis, Man, and Society*. 3rd ed. New
885 Brunswick, New Jersey: Rutgers University Press; 1952.
- 886 63. Kühnert D, Coscolla M, Brites D, Stucki D, Metcalfe J, Fenner L, et al. Tuberculosis
887 outbreak investigation using phylodynamic analysis. *Epidemics*. 2018 Dec 1;25:47–53.
- 888 64. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals
889 temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *PNAS*. 2013
890 Jan 2;110(1):228–33.
- 891 65. Boskova V, Stadler T, Magnus C. The influence of phylodynamic model specifications
892 on parameter estimates of the Zika virus epidemic. *Virus Evol* [Internet]. 2018 Jan 1
893 [cited 2018 Nov 2];4(1). Available from:
894 <https://academic.oup.com/ve/article/4/1/vex044/4829709>
- 895 66. Möller S, Plessis L du, Stadler T. Impact of the tree prior on estimating clock rates
896 during epidemic outbreaks. *PNAS*. 2018 Mar 28;201713314.
- 897 67. Jankute M, Nataraj V, Lee OY-C, Wu HHT, Ridell M, Garton NJ, et al. The role of
898 hydrophobicity in tuberculosis evolution and pathogenicity. *Scientific Reports*. 2017
899 May 2;7(1):1315.
- 900 68. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, et al.
901 Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex
902 indicates evolutionarily recent global dissemination. *PNAS*. 1997 Sep 2;94(18):9869–
903 74.
- 904 69. Gansauge M-T, Gerber T, Glocke I, Korlević P, Lippik L, Nagel S, et al. Single-stranded
905 DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids*
906 *Res*. 2017 Jun 2;45(10):e79–e79.
- 907 70. Andrades Valtueña A, Mittnik A, Key FM, Haak W, Allmäe R, Belinskij A, et al. The
908 Stone Age Plague and Its Persistence in Eurasia. *Current Biology* [Internet]. 2017 Nov
909 22 [cited 2017 Nov 30]; Available from:
910 <http://www.sciencedirect.com/science/article/pii/S0960982217313283>
- 911 71. Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren K-G, et al. Early
912 Divergent Strains of *Yersinia pestis* in Eurasia 5,000 Years Ago. *Cell*. 2015 Oct
913 22;163(3):571–82.
- 914 72. Spyrou MA, Tikhbatova RI, Wang C-C, Valtueña AA, Lankapalli AK, Kondrashin VV, et
915 al. Analysis of 3800-year-old *Yersinia pestis* genomes suggests Bronze Age origin for
916 bubonic plague. *Nature Communications*. 2018 Jun 8;9(1):2234.

- 917 73. Loosdrecht M van de, Bouzouggar A, Humphrey L, Posth C, Barton N, Aximu-Petri A,
918 et al. Pleistocene North African genomes link Near Eastern and sub-Saharan African
919 human populations. *Science*. 2018 May 4;360(6388):548–52.
- 920 74. Schuenemann VJ, Peltzer A, Welte B, van Pelt WP, Molak M, Wang C-C, et al. Ancient
921 Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in
922 post-Roman periods. *Nature Communications*. 2017 May 30;8:15694.
- 923 75. Meyer M, Kircher M. Illumina Sequencing Library Preparation for Highly Multiplexed
924 Target Capture and Sequencing. *Cold Spring Harbor Protocols*. 2010 Jun
925 1;2010(6):pdb.prot5448-pdb.prot5448.
- 926 76. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex
927 sequencing on the Illumina platform. *Nucleic Acids Research*. 2012 Jan 1;40(1):e3–e3.
- 928 77. Neukamm J, Peltzer A. Integrative-Transcriptomics/DamageProfiler [Internet]. 2018.
929 Available from: <http://doi.org/10.5281/zenodo.1291355>
- 930 78. Auwera GAV der, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, et
931 al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit
932 Best Practices Pipeline. *Current Protocols in Bioinformatics*. 2013 Oct 1;43(1):11.10.1-
933 11.10.33.
- 934 79. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
935 large phylogenies. *Bioinformatics*. 2014 May 1;30(9):1312–3.
- 936 80. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of
937 heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*
938 [Internet]. 2016 Jan 1 [cited 2017 Nov 30];2(1). Available from:
939 <https://academic.oup.com/ve/article/2/1/vew007/1753488>
- 940 81. Keane T, Creevey C, Pentony M, Naughton TJ, McInerney J. Assessment of methods
941 for amino acid matrix selection and their use on empirical data shows that ad hoc
942 assumptions for choice of matrix are not justified. *BMC evolutionary biology*.
943 2006;6:29–46.
- 944 82. Rambaut A, Suchard M, Xie W, Drummond AJ. Tracer. Institute of Evolutionary
945 Biology, University of Edinburgh; 2014.
- 946
- 947

948 ADDITIONAL FILES

949 **Additional File 1**

950 Format: Excel spreadsheet (.xlsx)

951 Title: Supplementary Tables

952 Description: Large tables of data contributing to the analyses presented in this paper,
953 including a taxon table showing assigned reads from all taxonomic levels represented in the
954 metagenomic LUND1 library (Table S1); full EAGER pipeline results for LUND1 shotgun
955 sequencing data when mapped to HG19 human reference genome and TB ancestor
956 genome, the non-UDG-treated enriched LUND1 data when mapped to the TB ancestor
957 genome, and the UDG-treated enriched LUND1 data when mapped to the TB ancestor
958 genome (Table S2); full EAGER pipeline results for negative controls processed with
959 LUND1, mapped to the reconstructed TB ancestor genome (Table S3); genomes included in
960 the full MTBC dataset, with respective publications, accession numbers, lineages, and dates
961 (when applicable) (Table S4); genomes included in the L4 dataset, with respective
962 publications, accession numbers, lineages, dates, and percentage of total SNPs called as
963 heterozygous (Table S5); sites excluded from the full MTBC dataset (Table S6); sites
964 excluded from the L4 dataset (Table S7); SnpEff annotation for derived alleles in LUND1
965 (Table S8).

966 **Additional File 2**

967 Format: Word document (.docx)

968 Title: Supplementary Information

969 Description: Detailed supplements to the RESULTS and METHODS sections,
970 including supplementary figures.

971