


RESEARCH

Open Access



# A seventeenth-century *Mycobacterium tuberculosis* genome supports a Neolithic emergence of the *Mycobacterium tuberculosis* complex

Susanna Sabin<sup>1</sup> , Alexander Herbig<sup>1</sup>, Åshild J. Vågane<sup>1,2</sup>, Torbjörn Ahlström<sup>3</sup>, Gracijela Bozovic<sup>4</sup>, Caroline Arcini<sup>5</sup>, Denise Kühnert<sup>6\*</sup> and Kirsten I. Bos<sup>1\*</sup>

\* Correspondence: [kuehnert@shh.mpg.de](mailto:kuehnert@shh.mpg.de); [bos@shh.mpg.de](mailto:bos@shh.mpg.de)

<sup>6</sup>Transmission, Infection, Diversification & Evolution Group, Max Planck Institute for the Science of Human History, 07745 Jena, Germany

<sup>1</sup>Department of Archaeogenetics, Max Planck Institute for the Science of Human History, 07745 Jena, Germany

Full list of author information is available at the end of the article

## Abstract

**Background:** Although tuberculosis accounts for the highest mortality from a bacterial infection on a global scale, questions persist regarding its origin. One hypothesis based on modern *Mycobacterium tuberculosis* complex (MTBC) genomes suggests their most recent common ancestor followed human migrations out of Africa approximately 70,000 years before present. However, studies using ancient genomes as calibration points have yielded much younger dates of less than 6000 years. Here, we aim to address this discrepancy through the analysis of the highest-coverage and highest-quality ancient MTBC genome available to date, reconstructed from a calcified lung nodule of Bishop Peder Winstrup of Lund (b. 1605–d. 1679).

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

**Conclusions:** The results of our metagenomic analysis demonstrate the unique preservation environment calcified nodules provide for DNA. Importantly, we estimate a most recent common ancestor date for the MTBC of between 2190 and 4501 before present and for Lineage 4 of between 929 and 2084 before present using multiple models, confirming a Neolithic emergence for the MTBC.

**Keywords:** Tuberculosis, Ancient DNA, *Mycobacterium tuberculosis*, Molecular dating, Metagenomics



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Tuberculosis, caused by organisms in the *Mycobacterium tuberculosis* complex (MTBC), has taken on renewed relevance and urgency in the twenty-first century due to its global distribution, its high morbidity, and the rise of antibiotic-resistant strains [1]. The difficulty in disease management and treatment, combined with the massive reservoir the pathogen maintains in human populations through latent infection [2], makes tuberculosis a pressing public health challenge. Despite this, controversy exists regarding the history of the relationship between members of the MTBC and their human hosts.

Existing literature suggests two estimates for the time to the most recent common ancestor (tMRCA) for the MTBC based on the application of Bayesian molecular dating to genome-wide *Mycobacterium tuberculosis* data. One estimate suggests the extant MTBC emerged through a bottleneck approximately 70,000 years ago, coincident with major migrations of humans out of Africa [3]. This estimate was reached using a large global dataset of exclusively modern *M. tuberculosis* genomes, with internal nodes of the MTBC calibrated by extrapolated dates for major human migrations [3]. This estimate relied on congruence between the topology of the MTBC and human mitochondrial phylogenies, but this congruence does not extend to human Y chromosome phylogeographic structure [4]. As an alternative approach, the first publication of ancient MTBC genomes utilized radiocarbon dates as direct calibration points to infer mutation rates and yielded an MRCA date for the complex of less than 6000 years [5]. This younger emergence was later supported by mutation rates estimated within the pervasive Lineage 4 (L4) of the MTBC, using four *M. tuberculosis* genomes from the late eighteenth and early nineteenth centuries [6].

Despite the agreement in studies that have relied on ancient DNA calibration so far, dating of the MTBC emergence remains controversial. The young age suggested by these works cannot account for purported detection of MTBC DNA in archeological material that predates the tMRCA estimate (e.g., Baker et al. [7]; Hershkovitz et al. [8]; Masson et al. [9]; Rothschild et al. [10]), the authenticity of which has been challenged [11]. Furthermore, constancy in mutation rates of the MTBC has been questioned on account of observed rate variation in modern lineages, combined with the unquantified effects of latency [12]. The ancient genomes presented by Bos and colleagues, though isolated from human remains, were most closely related to *Mycobacterium pinnipedii*, a lineage of the MTBC currently associated with infections in seals and sea lions [5]. Given our unfamiliarity with the demographic history of tuberculosis in sea mammal populations [13], identical substitution rates between the pinniped lineage and human-adapted lineages of the MTBC cannot be assumed. Additionally, estimates of genetic diversity in MTBC strains from archeological specimens can be difficult given their similarities to environmental mycobacterial DNA from the depositional context, which increase the risk of false positive genetic characterization [14]. Though the ancient genomes published by Kay and colleagues belonged to human-adapted lineages of the MTBC, and the confounding environmental signals were significantly reduced by their funerary context in crypts, two of the four genomes used for molecular dating were derived from mixed-strain infections [6]. By necessity, diversity derived in each genome would have to be ignored for them to be computationally

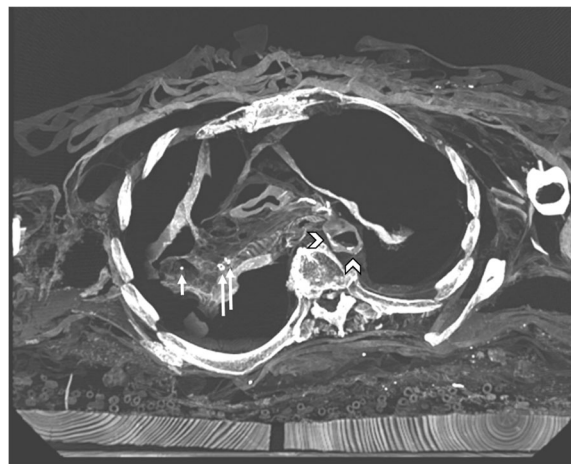
distinguished [6]. Though ancient DNA is a valuable tool for answering the question of when the MTBC emerged, the available ancient data remains sparse and subject to case-by-case challenges.

Here, we offer a higher resolution temporal estimate for the emergence of the MTBC and L4 using multiple Bayesian models of varying complexity through the analysis of a high-coverage seventeenth-century *M. tuberculosis* genome extracted from a calcified lung nodule. Removed from naturally mummified remains, the nodule provided an excellent preservation environment for the pathogen, and exhibited minimal infiltration by exogenous bacteria. The nodule and surrounding lung tissue also showed exceptional preservation of host DNA, thus showing promise for this tissue type in ancient DNA investigations.

## Results

### Pathogen identification

Computed tomography (CT) scans of the mummified remains of Bishop Peder Winstrup of Lund, Sweden revealed a calcified granuloma a few millimeters (mm) in size in the collapsed right lung together with two ~5 mm calcifications in the right hilum (Fig. 1). Primary tuberculosis causes parenchymal changes and ipsilateral hilar lymphadenopathy that is more common on the right side [15]. Upon resolution, it can leave a parenchymal scar, a small calcified granuloma (Ghon focus), and calcified hilar nodes, which are together called a Ranke complex. In imaging, this complex is suggestive of previous tuberculosis infection, although histoplasmosis can have the same appearance [16]. Histoplasmosis, however, is very rare in Scandinavia and is more often seen in other parts of the world (e.g., the Americas) [17]. The imaging findings were therefore considered to result from previous primary tuberculosis. One of the calcified hilar nodes was extracted from the remains during video-assisted thoracoscopic surgery, guided by fluoroscopy. The extracted material was further subsampled for genetic



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

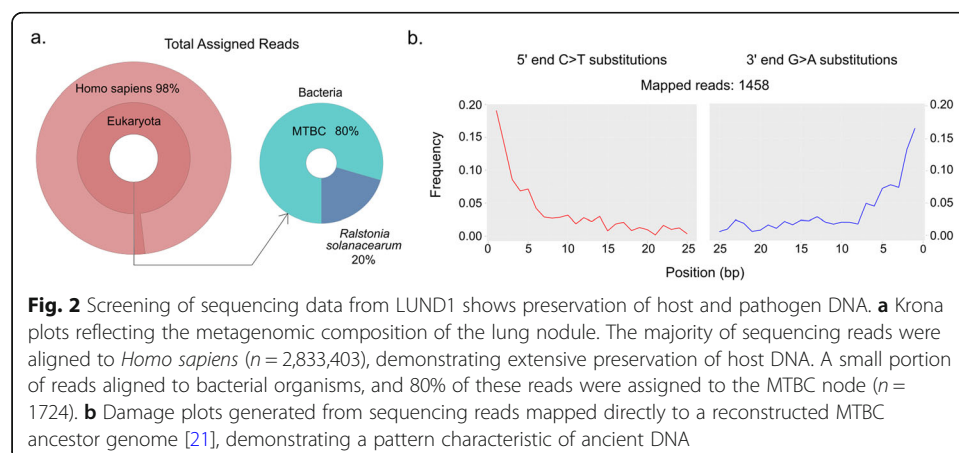
analysis. DNA was extracted from the nodule and accompanying lung tissue using protocols optimized for the recovery of ancient, chemically degraded, fragmentary genetic material [18]. The library (LUND1) was shotgun sequenced to a depth of approximately 3.7 million reads.

Adapter-clipped and base quality-filtered reads were taxonomically binned with MALT [19] against the full NCBI Nucleotide database (“nt,” April 2016). In this process, 3,515,715 reads, or 95% of the metagenomic reads, could be assigned to taxa contained within the database. Visual analysis of the metagenomic profile in MEGAN6 [20] revealed the majority of these reads, 2,833,403 or 81%, were assigned to *Homo sapiens*. A further 1724 reads were assigned to the *Mycobacterium tuberculosis* complex (MTBC) node. Importantly, no other taxa in the genus *Mycobacterium* were identified, and the only other identified bacterial taxon was *Ralstonia solanacearum* (Fig. 2a), a soil-dwelling plant pathogen frequently identified in metagenomic profiles of archaeological samples [22, 23] (Additional File 1).

Pre-processed reads were mapped to both the hg19 human reference genome and a reconstructed MTBC ancestor (TB ancestor) [21] using BWA as implemented in the Efficient Ancient Genome Reconstruction (EAGER) pipeline [24]. Reads aligned to hg19 with direct mapping constituted an impressive 88% of the total sequencing data (Additional File 2). Human mitochondrial contamination was extremely low, estimated at only 1–3% using Schmutzi [25] (Additional File 3). Reads were also mapped to the TB ancestor (Table 1). After map quality filtering and read de-duplication, 1458 reads, or 0.045% of the total sequencing data, aligned to the reference (Table 1) and exhibited cytosine-to-thymine damage patterns indicative of authentic ancient DNA (Fig. 2b) [26, 27]. Qualitative preservation of the tuberculosis DNA was slightly better than that of the human DNA, as damage was greater in the latter (Additional File 2). Laboratory-based contamination, as monitored by negative controls during the extraction and library preparation processes, could be ruled out as the source of this DNA (Additional File 4).

### Genomic enrichment and reconstruction

Due to the clear but low-abundance MTBC signal, a uracil DNA glycosylase (UDG) library was constructed to remove DNA lesions caused by hydrolytic deamination of



**Table 1** Mapping statistics for LUND1 libraries

Pre/post capture	Library treatment	Processed reads pre-mapping (n)	Unique mapped reads, quality-filtered (n)	Endogenous DNA (%)	Mean fold coverage	Mean fragment length (bp)	GC content (%)
Pre-capture	Non-UDG	3,696,712	1458	0.045	0.018	54.31	63.89
Post-capture	UDG	59,091,507	9,482,901	45.652	141.5062	65.83	62.96

A comparison of the mapping statistics for the non-UDG screening library and UDG-treated MTBC enriched library of LUND1 when aligned to the MTBC ancestor genome [21]. For full EAGER output, see Additional File 2

cytosine residues [28] and enriched with an in-solution capture [29, 30] designed to target genome-wide data representing the full diversity of the MTBC (see the “Methods” section). The capture probes are based on a reconstructed TB ancestor genome [21]. The enriched library was sequenced using a paired-end, 150-cycle Illumina sequencing kit to obtain a full fragment-length distribution (Fig. S1 in Additional File 3). The resulting sequencing data was then aligned to the hypothetical TB ancestor genome [21], and the mapping statistics were compared with those from the screening data to assess enrichment (Table 1). Enrichment increased the proportion of endogenous MTBC DNA content by three orders of magnitude, from 0.045 to 45.652%, and deep sequencing yielded genome-wide data at an average coverage of approximately 141.5-fold. The mapped reads have an average fragment length of ~ 66 base pairs (Table 1).

We further evaluated the quality of the reconstructed genome by quantifying the amount of heterozygous positions (see the “Methods” section). Derived alleles represented by 10–90% of the reads covering a given position with five or more reads of coverage were counted. Only 24 heterozygous sites were counted across all variant positions in LUND1. As a comparison, the other high-coverage (~ 125 fold) ancient genome included here—body92 from Kay et al. [6]—contained 70 heterozygous positions.

### Phylogeny and dating

Preliminary phylogenetic analysis using neighbor joining (Figs. S2 and S3 in Additional File 3), maximum likelihood (Figs. S4 and S5 in Additional File 3), and maximum parsimony trees (Figs. S6 and S7 in Additional File 3) indicated that LUND1 groups within the L4 strain diversity of the MTBC, and more specifically, within the L4.10/PGG3 sublineage. This sublineage was recently defined by Stucki and colleagues as the clade containing L4.7, L4.8, and L4.9 [31] according to the widely accepted Coll nomenclature [32]. Following this, we constructed two datasets to support molecular dating of the full MTBC (Additional File 5) and L4 of the MTBC (Additional File 6).

The dataset reflecting extant diversity of the MTBC was compiled as reported elsewhere [5], with six ancient genomes as calibration points. These included LUND1; two additional ancient genomes, body80 and body92, extracted from late 18th and early nineteenth century Hungarian mummies [6]; and three human-isolated *Mycobacterium pinnipedii* strains from Peru [5], encompassing all available ancient *M. tuberculosis* genomes with sufficient coverage to call SNPs confidently after stringent mapping with BWA [33] (see the “Methods” section; Additional File 5). *Mycobacterium canettii* was used as an outgroup. In generating an alignment of variant positions in this dataset, we

excluded repetitive regions and regions at risk of cross-mapping with other organisms as done previously [5], as well as potentially imported sites from recombination events, which were identified using ClonalFrameML [34] (Additional File 7). We chose to exclude these potential recombinant sites despite *M. tuberculosis* being generally recognized as a largely clonal organism with no recombination or horizontal gene transfer, as these phenomena have been found to occur in *M. canettii* [35, 36]. Only twenty-three variant sites were lost from the full MTBC alignment as potential imports. We called a total of 42,856 variable positions in the dataset as aligned to the TB ancestor genome. After incompletely represented sites were excluded, 11,716 were carried forward for downstream analysis. Prior to performing the Bayesian molecular dating analysis, we assessed the dataset for clock-like structure with TempEst ( $R^2 = 0.273$ ; see the “Methods” section; Fig. S8 in Additional File 3).

To explore the impact of the selected tree prior and clock model, we ran multiple variations of models as available for use in BEAST2 [37]. We first used both a strict and a relaxed clock model together with a constant coalescent model (CC+strict, CC+UCLD). We found there to be minimal difference between the inferred rates estimated by the two models. This finding, in addition to the low rate variance estimated in all models, suggests there is little rate variation between known branches of the MTBC. Nevertheless, the relaxed clock appeared to have a slightly better performance (Table 2). To experiment with models that allowed for dynamic populations, we applied a Bayesian skyline coalescent (SKY+UCLD) and birth-death skyline prior (BDSKY+UCLD) combined with a relaxed clock model. In the BDSKY+UCLD model, the tree was conditioned on the root. In a prior study, Kühnert and colleagues used birth-death tree priors to investigate two modern tuberculosis outbreaks [38]. To our knowledge, this study is the first to use a birth-death tree prior to infer evolutionary dynamics of the MTBC while using ancient data for tip calibration. The BDSKY+UCLD model had the highest marginal likelihood value of all models applied to this dataset (Table 2).

A calibrated maximum clade credibility (MCC) tree was generated for the BDSKY+UCLD model, with 3258 years before present (BP) (95% highest posterior density [95% HPD] interval, 2190–4501 BP) as an estimated date of emergence for the MTBC (Fig. 3a). Tree topology agrees with previously presented phylogenetic analyses of the full MTBC [3, 5, 39]. To test the meaningfulness of our ancient tip calibrations,

**Table 2** Model comparison for full MTBC dataset

Model	Marginal likelihood	Mean rate (95% HPD)	Mean rate variance (95% HPD)	Mean tree height (95% HPD)
BDSKY+UCLD	-6125044.47176458	1.4488E-8 (9.4606E-9, 1.9632E-8)	1.6881E-17 (5.4855E-18, 3.069E-17)	3258.0478 (2189.5235, 4501.1384)
CC+UCLD	-6126017.15694528	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	-6127733.35000634	1.2944E-8 (8.6149E-9, 1.7342E-8)	1.3423E-17 (4.848E-18, 2.3869E-17)	3650.4222 (2472.6434, 4992.0277)
CC+strict	-6125541.68118691	1.1573E-8 (8.6397E-9, 1.4509E-8)	NA	4453.1162 (3330.1516, 5619.3974)

Marginal likelihood and parameter estimates from four models applied to the full MTBC dataset: constant coalescent with uncorrelated lognormal clock (CC+UCLD), constant coalescent with strict clock (CC+strict), Bayesian skyline coalescent with uncorrelated lognormal clock (SKY+UCLD), and birth-death skyline with uncorrelated lognormal clock (BDSKY+UCLD). Marginal likelihoods obtained through path sampling (see the “Methods” section)





excluding positions unique to the L2 outgroup. Only fifteen variant sites were lost from the L4 dataset alignment. After sites missing from any alignment in the dataset were excluded from downstream analysis, 10,009 SNPs remained for phylogenetic inference. A total of 810 SNPs were identified in LUND1, of which 126 were unique to this genome. A SNP effect analysis [42] was subsequently performed on these derived positions (Additional File 3; Additional File 9). We also assessed the L4 dataset for clock-like structure with TempEst ( $R^2 = 0.113$ ; see the “Methods” section; Fig. S9 in Additional File 3).

We applied the same models as described above for the full MTBC dataset, with the addition of a birth-death skyline model conditioned on the origin of the root (BDSKY+UCLD+origin). All mean tree heights are within 250 years of each other and the 95% HPD intervals largely overlap. The BDSKY+UCLD and BDSKY+UCLD+origin models show the highest marginal likelihood values after stepping stone sampling. We employed the BDSKY+UCLD+origin model to determine if the estimated origin of the L4 dataset agreed with the tree height estimates for the full MTBC dataset. Intriguingly, the estimated origin parameter (Table 3), or the ancestor of the tree root, largely overlaps with the 95% HPD range for MTBC tree height as seen in Table 2.

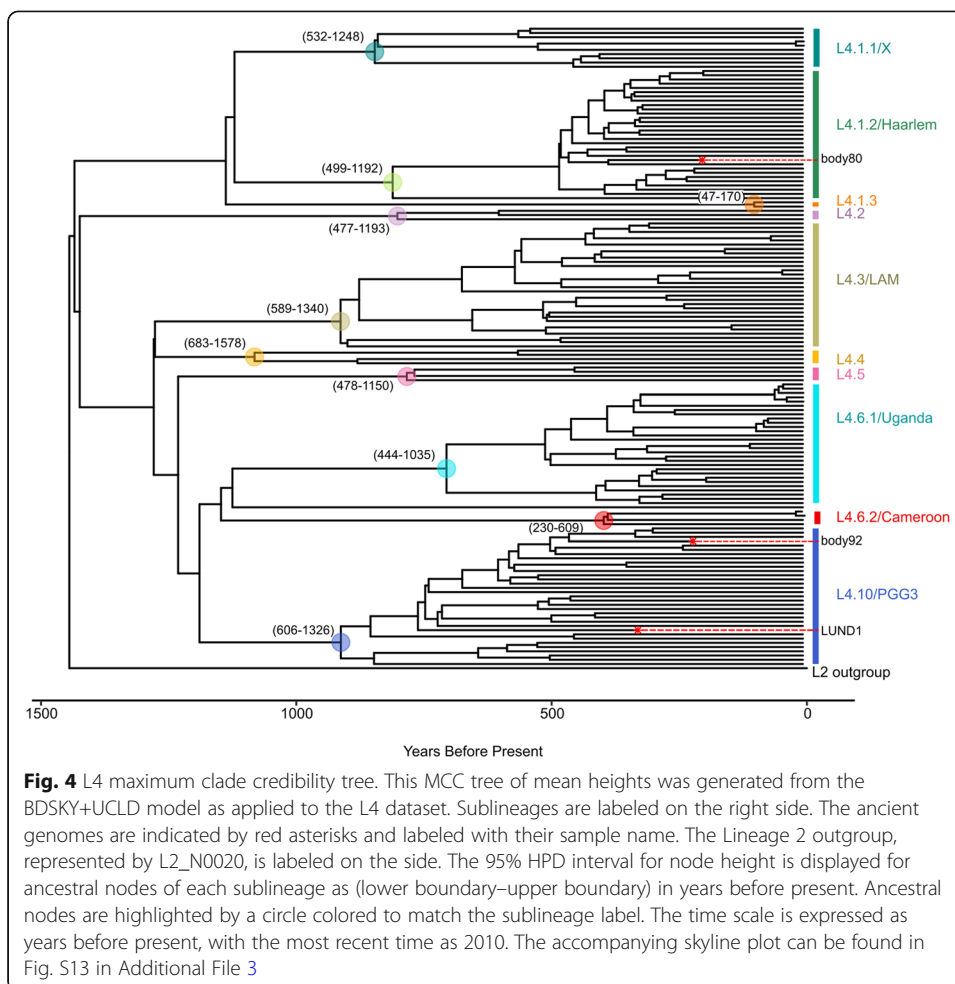
A calibrated MCC tree (Fig. 4) was generated based on the BDSKY+UCLD model for the L4 dataset. This model yielded an estimated date of emergence for L4 of 1445 BP (95% HPD, 929–2084 BP). The tree reflects the ten-sublineage topology presented by Stucki and colleagues [31], with LUND1 grouping with the L4.10/PGG3 sublineage. Due to the relatively low  $R^2$  value for the relationship between sampling time and root-to-tip distance as calculated using TempEst, we also performed a date randomization test of the L4 BDSKY+UCLD model, in which we shuffled the sampling dates randomly among all genomes [40, 41]. We performed ten randomizations and compared the resulting clock rate estimates with that of the BDSKY+UCLD model with the true sampling dates (Table 3). Nine out of ten randomizations fulfilled the more stringent criterion ii, exhibiting no overlap between their 95% HPD intervals and that of the original

**Table 3** Model comparison for L4 dataset

Model	Marginal likelihood	Mean rate (95% HPD)	Mean rate variance (95% HPD)	Mean tree height (95% HPD)	Origin (BDSKY only)
BDSKY+UCLD	− 6033864.2003	3.1885E−8 (1.9488E−8, 4.4007E−8)	4.991E−17 (1.0674E−17, 8.9835E−17)	1444.5416 (929.3966, 2083.7636)	NA
BDSKY+UCLD +origin	− 60327945.1483	3.4761E−8 (2.447E−8, 4.5029E−8)	5.5123E−17 (1.9718E−17, 9.4555E−17)	1319.2463 (952.8702, 1761.4382)	2310.916 (1165.2155, 3372.9253)
CC+UCLD	− 6043356.1504	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	− 6034698.3620	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	− 6034091.5119	2.9299E−8 (2.2173E−8, 3.6637E−8)	NA	1567.544 (1186.1186, 1978.6488)	NA

Selected parameter estimates from five models applied to the Lineage 4 dataset: constant coalescent with uncorrelated lognormal clock (CC+UCLD), constant coalescent with strict clock (CC+strict), Bayesian skyline coalescent with uncorrelated lognormal clock (SKY+UCLD), birth-death skyline with uncorrelated lognormal clock and tree conditioned on the root (BDSKY+UCLD), and birth-death skyline with uncorrelated lognormal clock with origin parameter estimate (BDSKY+UCLD+origin). Marginal likelihoods obtained through path sampling (see the “Methods” section)





(Additional File 11; Fig. S12 in Additional File 3). All ten randomizations satisfied criterion i (i.e., yielded a mean rate estimate that fell outside the 95% HPD interval of the rate from the model using true temporal values).

## Discussion

The increasing number of ancient *Mycobacterium tuberculosis* genomes is steadily reducing the uncertainty of molecular dating estimates for the emergence of the MTBC. Here, using the ancient data available to date, we directly calibrate the MTBC time tree and confirm that known diversity within the complex is derived from a common ancestor that existed ~2000–6000 years before present (Fig. 3; Table 2) [5, 6]. Our results support the hypothesis that the MTBC emerged during the Neolithic, and not before. The Neolithic revolution generally refers to the worldwide transition in lifestyle and subsistence from more mobile, foraging economies to more sedentary, agricultural economies made possible by the domestication of plants and animals. The period during which it occurred varies between regions. In Africa, where the MTBC is thought to have originated [3, 43–45], the onset of these cultural changes, and animal domestication in particular, appears to have its focus around ~3000 BCE, or 5000 BP, across multiple regions [46]. The estimates presented here place the emergence of tuberculosis

amidst the suite of human health impacts that took place as a consequence of the Neolithic lifestyle changes often referred to collectively as the first epidemiological transition [47, 48].

Tuberculosis has left testaments to its history as a human pathogen in the archeological record [49], where some skeletal analyses have been interpreted to suggest tuberculosis in human and animal remains pre-dating the upper 95% HPD boundary for the MTBC tMRCA presented here [7, 8, 10, 50–54]. However, it is important to explore the evolutionary history of the MTBC through molecular data. Furthermore, it is crucial to base molecular dating estimates on datasets that include ancient genomes, which expand the temporal sampling window and provide data from the pre-antibiotic era. Numerous studies have found long-term nucleotide substitution rate estimates in eukaryotes and viruses to be dependent on the temporal breadth of the sampling window, and it is reasonable to assume the same principle applies to bacteria [55–60]. Additionally, rate variation over time and between lineages, which may arise due to changing evolutionary dynamics such as climate and host biology, can impact the constancy of the molecular clock [58, 59]. Though models have been developed to accommodate uncertainty regarding these dynamics [61], temporally structured populations can provide evidence and context for these phenomena over time and can aid researchers in refining models appropriate for the taxon in question [60]. Though we did not identify substantial rate variation within either the MTBC or L4 trees (Figs. S14 and S15 in Additional File 3), it is important that we draw these observations from temporally structured datasets and continue to do so in the future.

In addition to our tMRCA estimate for the MTBC, we present one for L4, which is among the most globally dominant lineages in the complex [31, 62]. Our analyses yielded tMRCA dates between ~900 and 2500 years before present, as extrapolated from the 95% HPD intervals of all models (Table 3), with mean dates spanning from 320 to 691 CE. These results are strikingly similar to those found in two prior publications and support the idea proposed by Kay and colleagues that L4 may have emerged during the late Roman period [5, 6]. However, there exist discrepancies between different estimates for the age of this lineage in available literature that overlap with the upper [63] and lower [62] edges of the 95% HPD intervals reported here. In addition, recent phylogeographic analyses of the MTBC and its lineages had ambiguous results for L4, with the internal nodes being assigned to either African or European origins depending on the study or different dataset structures used within the same study [62, 63]. This finding indicates a close relationship between ancestral L4 strains in Europe and Africa [62, 63]. Stucki and colleagues delineated L4 into two groups based on the extent of their geographic distribution: globally distributed “generalist” sublineages and highly local “specialist” sublineages that do not appear outside a restricted geographical niche [31]. Thus far, the “specialist” sublineages are found regionally on the African continent. A clear phylogenetic relationship explaining the distinction between geographically expansive and limited strains has not been established. Specifically, LUND1 falls within the globally distributed, “generalist” L4.10/PGG3 sublineage that shares a clade with two “specialist” sublineages: L4.6.1/Uganda and L4.6.2/Cameroon (Fig. 4) [31]. Elucidating the phenomenon that separated L4.10/PGG3 and the L4.6 lineages could offer relevant clues about the evolutionary relationship between specific populations of MTBC organisms and specific populations of humans by selection or genetic

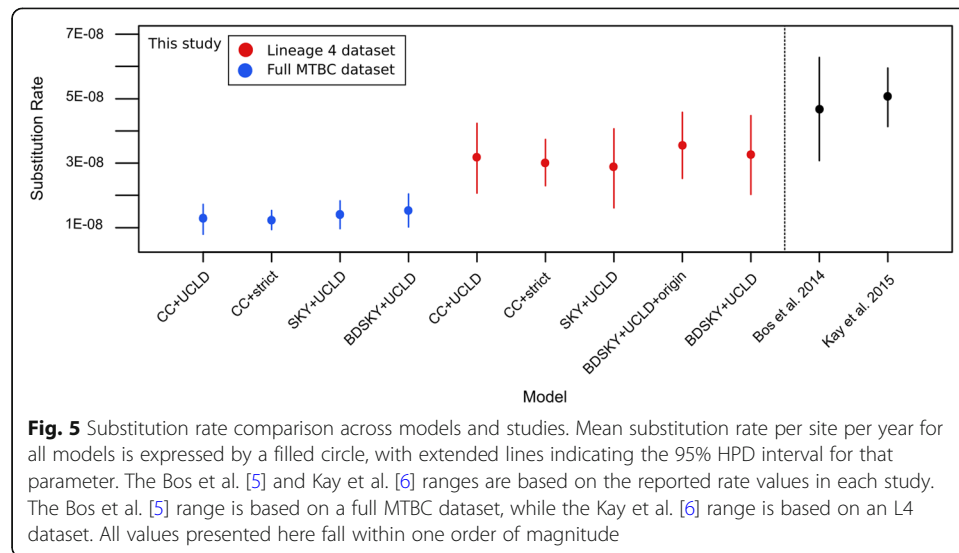
drift discussed elsewhere [44, 64]. Assuming modern L4 diversity in Africa was driven by exchanges between Europe and Africa [62, 63], why do we not see the L4.6 lineages more frequently in European populations as we do their sister clade? The current discrepancies over the age and geographic origin of L4 make interpretations of existing data unreliable for questions of such specificity and complexity at this time. These discrepancies could be due to differences in genome selection, SNP selection, and/or model selection and parameterization. It is unlikely we will gain clarity until more diverse, high-quality ancient L4 genomes are generated, creating a more temporally and geographically structured dataset.

Going deeper into comparisons between the results presented here and those from prior studies, mutation rate estimates in the L4 and full MTBC analyses were lower than previous estimates for comparable datasets, but within the same order of magnitude, with all mean and median estimates ranging between  $1\text{E}-8$  and  $5\text{E}-8$  [5, 6] (Table 2). Nucleotide substitution rates inferred based on modern tuberculosis data are close to but slightly higher than those based on ancient calibration, with multiple studies finding rates of approximately  $1\text{E}-7$  substitutions per site per year [4, 65]. Despite a strict clock model having been rejected by the MEGA-CC molecular clock test [66] for both the L4 and full MTBC datasets, the clock rate variation estimates do not surpass  $9\text{E}-17$  in any model. Additionally, there is little difference between the clock rates estimated in the L4 and full MTBC datasets suggesting the rate of evolution in L4 does not meaningfully differ from that of the full complex (Tables 2 and 3; Fig. 5).

Importantly, we explored our data through multiple models, including birth-death tree priors. In our opinion, these models offer more robust parameterization options for heterochronous datasets that are unevenly distributed over time, such as those presented here, by allowing for uneven sampling proportions across different time intervals of the tree [67]. Recent studies have demonstrated the importance of selecting appropriate tree priors for the population under investigation, as well as the differences between birth-death and coalescent tree priors [68, 69]. It is notable that the estimates reported here roughly agree across multiple demographic and clock models implemented in BEAST2. The estimate of the origin height for the L4 dataset as calculated with the birth-death Skyline model overlaps with the 95% HPD intervals for the tree height estimates across models in the full MTBC dataset.

In addition to confirming the findings of prior publications, this study contributes a high-coverage, contamination-free, and securely dated ancient *M. tuberculosis* genome for future dating efforts, which may include more ancient data or more realistic models. Much of this quality likely comes from the unique preservation environment of the calcified nodule. In the case of tuberculosis, such nodules form from host immunological responses in the waning period of an active pulmonary infection [70] and remain in lung tissue, characterizing the latent form of the disease. Host immune cells were likely responsible for the dominant signal of human DNA in the LUND1 metagenomic screening library (Fig. 1, Supplementary Table 2 in Additional File 1). Similar levels of preservation have been observed through analyses of ancient nodules yielding *Brucella* [71] and urogenital bacterial infections [72], with pathogen preservation surpassing what we report here.

LUND1 avoided multiple quality-related problems often encountered in the identification and reconstruction of ancient genetic data from the MTBC. The genome is of



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

While the high quality and securely dated ancient genome presented here offered advantages in a molecular dating approach, there are caveats to the results of this study. First, this analysis excludes diversity within *M. canettii*—a bacterium that can cause pulmonary tuberculosis—from the MTBC dataset, and as such, our estimate does not preclude the possibility of a closely related ancestor having caused infections indistinguishable from tuberculosis in humans before 6000 BP. The inferred tMRCA could be restricted to a lineage that survived an evolutionary bottleneck or selective sweep, possibly connected to its virulence in humans as suggested elsewhere, albeit as a considerably more ancient event [45, 73, 74]. It is possible there were pathogenic sister lineages to the MTBC that existed prior to this reduction in diversity and are not represented by extant MTBC diversity. Additionally, despite the use of ancient data, our temporal sampling window is still narrow given the estimated age of the MTBC and L4. For the MTBC dataset no samples pre-date 1000 years before present, and for L4, no samples predate 350 years before present. It could be argued the ancient L4 genomes available to date represent samples taken in the midst of an epidemic—namely, the “White Plague” of tuberculosis, which afflicted Europe between the seventeenth and nineteenth

centuries [75]. For a slow-evolving bacterial pathogen like tuberculosis, it is possible our sampling window of ancient genomes is subject to the very issue they are meant to alleviate: the time dependency of molecular clocks [55, 57–59]. The genomes sampled from pre-contact Peruvian remains do not derive from a known epidemic period in history and add temporal spread to our MTBC dataset, but also belong to a clade of animal-associated strains (*M. pinnipedii*) that may have been subject to dramatically different evolutionary pressures compared to the human-associated lineages of the complex due to differing host biology and population dynamics. However, our use of a relaxed clock model allowed for the estimation and accommodation of variable rates across different branches of the complex. We do not see evidence for divergent substitution rates among the branches leading to the Peruvian *M. pinnipedii* strains (Fig. S14 in Additional File 3). On a related matter, we may be missing diversity for some lineages (e.g., L6, L7, animal lineages) for which whole genome data is sparse. The available ancient MTBC genomes also suffer from a lack of lineage diversity, with only pinniped strains and L4 represented. We furthermore qualify our BDSKY results by acknowledging our models required the specification of priors for the  $\rho$  parameter (the sampling proportion of the total population at discrete time points). We chose  $\rho$  priors (see Additional File 3) assuming that our modern genomes represented a greater sampling proportion of the total contemporaneous MTBC and L4 populations than our ancient genomes. This assumption alone made this parameterization less arbitrary than the assumptions inherent in the coalescent-based methods that have been utilized in the past for similar time-sampled analyses of the MTBC and other pathogens, which assume random sampling at uniform rates across all time periods. We also acknowledge that skyline models assume panmictic populations, and the datasets presented here do contain spatial subdivision, which may bias estimates regarding population dynamics. However, this aspect of our datasets is unlikely to bias our molecular clock estimates. As stated above, the agreement of multiple models to reach similar dates for the tMRCA of the MTBC and L4 reinforces our support of the hypothesis that the most recent common ancestor of the MTBC diversity we are aware of today emerged during the Neolithic.

Filling the MTBC time tree with more ancient genomes from diverse time periods, locations, and lineages would have the potential to address the limitations listed above. The most informative data would (a) derive from an Old World context (i.e., Europe, Asia, or Africa) pre-dating the White Plague in Europe or (b) come from any geographical location or pre-modern time period, but belong to one of the MTBC lineages not yet represented by ancient data. An ideal data point, which would clarify many open questions and seeming contradictions related to the evolutionary history of the MTBC, would derive from Africa, the inferred home of the MTBC ancestor [3, 43–45], and pre-date 2000 years before present. A genome of this age would test the lower boundaries of the 95% HPD tree height intervals estimated in the full MTBC models presented here. Until recently, it would have been considered unrealistic to expect such data to be generated from that time period and location. Innovations and improvements in ancient DNA retrieval and enrichment methods, however, have brought this expectation firmly into the realm of the possible [30, 76]. Ancient bacterial pathogen genomes have now been retrieved from remains from up to 5000 years before present [77–79] and recent studies have reported the recovery of human genomes from up to 15,000-year-old remains from North Africa [80, 81].

## Conclusions

Here, we offer confirmation that the extant MTBC, and all available ancient MTBC genomes, stem from a common ancestor that existed a maximum of 6000 years before present. Many open questions remain, however, regarding the evolutionary history of the MTBC and its constituent lineages, as well as the role of tuberculosis in human history. Elucidating these questions is an iterative process, and progress will include the generation of diverse ancient *M. tuberculosis* genomes, and the refinement and improved parameterization of Bayesian models that reflect the realities of MTBC (and other organisms') population dynamics and sampling frequencies over time. To aid in future attempts to answer these questions, this study provides an ancient MTBC genome of impeccable quality and explores the first steps in applying birth-death population models to modern and ancient TB data.

## Methods

### Lung nodule identification

The paleopathological investigation of the body of Winstrup is based on extensive CT scan examinations with imaging of the mummy and its bedding performed with a Siemens Somatom Definition Flash, 128 slice at the Imaging Department of Lund University Hospital. Ocular inspection of the body other than of the head and hands was not feasible, since Winstrup was buried in his episcopal robes and underneath the body was wrapped in linen strips. The velvet cap and the leather gloves were removed during the investigation. The body was naturally mummified and appeared to be well preserved with several internal organs identified.

The imaging was quite revealing. The intracranial content was lost with remains of the brain in the posterior skull base. Further, the dental status was poor with several teeth in the upper jaw affected by severe attrition, caries, and signs of tooth decay, as well as the absence of all teeth in the lower jaw. Most of the shed teeth were represented by closed alveoli, indicating antemortem tooth loss. Along with the investigation of the bedding, a small sack made of fabric was found behind the right elbow containing five teeth: two incisors, two premolars, and one molar. The teeth in the bag complemented the remaining teeth in the upper jaw. It is feasible that the teeth belonged to Winstrup and were shed several years before he died. A fetus approximately 5 months of age was also found in the bedding, underneath his feet.

Both lungs were preserved but collapsed with findings of a small parenchymal calcification and two ~ 5 mm calcifications in the right hilum (Fig. 1). The assessment was that these could constitute a Ranke complex, suggestive of previous primary tuberculosis [70]. A laparoscopy was performed at the Lund University Hospital in a clinical environment whereby the nodules were retrieved. Furthermore, several calcifications were also found in the aorta and the coronary arteries, suggesting the presence of atherosclerosis. The stomach, liver, and gall bladder were preserved, and several small gallstones were observed. The spleen could be identified but not the kidneys. The intestines were there, however, collapsed except for the rectum that contained several large pieces of concernments. The bladder and the prostate could not be recognized.

The skeleton showed several pathological changes. Findings on the vertebrae consistent with DISH (diffuse idiopathic skeletal hyperostosis) were present in the thoracic



and the lumbar spine. Reduction of the joint space in both hip joints and the left knee joint indicate that Winstrup was affected by osteoarthritis. No signs of gout or osteological tuberculosis (i.e., Pott's disease) were found.

Neither written sources nor the modern examination of the body of Winstrup reveal the immediate cause of death. However, it is known that he was bedridden for at least 2 years preceding his death. Historical records indicate that gallstones caused him problems while traveling to his different parishes. Additionally, he was known to have suffered from tuberculosis as a child, which may have recurred in his old age.

### **Sampling and extraction**

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

### **Library preparation and shotgun screening sequencing**

Double-stranded Illumina libraries were constructed according to an established protocol with some modifications [82]. Overhangs of DNA fragments were blunt-end repaired in a 50 µL reaction including 10 µL of the LUND1 extract, 21.6 µL of H<sub>2</sub>O, 5 µL of NEB Buffer 2 (New England Biolabs), 2 µL dNTP mix (2.5 mM), 4 µL BSA (10 mg/ml), 5 µL ATP (10 mM), 2 µL T4 polynucleotide kinase, and 0.4 µL T4 polymerase, then purified and eluted in 18 µL TET. Illumina adapters were ligated to the blunt-end fragments in a reaction with 20 µL Quick Ligase Buffer, 1 µL of adapter mix (0.25 µM), and 1 µL of Quick Ligase. Purification of the blunt-end repair and adapter ligation steps was performed using MinElute columns (Qiagen). Adapter fill-in was performed in a 40-µL reaction including 20 µL adapter ligation eluate, 12 µL H<sub>2</sub>O, 4 µL Thermopol buffer, 2 µL dNTP mix (2.5 mM), and 2 µL Bst polymerase. After the reaction was incubated at 37 °C for 20 min, the enzyme was heat deactivated with a 20-min incubation at 80 °C. Four library blanks were processed alongside LUND1 to control for reagent/laboratory contamination. The library was quantified using a real-time qPCR assay (Lightcycler 480 Roche) with the universal Illumina adapter sequences IS7 and IS8 as targets. Following this step, the library was double indexed [83] with a unique pair of indices over two 100 µL reactions using 19 µL of template, 63.5 µL of H<sub>2</sub>O, 10 µL PfuTurbo buffer, 1 µL

PfuTurbo (Agilent), 1  $\mu$ l dNTP mix (25 mM), 1.5  $\mu$ l BSA (10 mg/ml), and 2  $\mu$ l of each indexing primer (10  $\mu$ M). The master mix was prepared in a pre-PCR clean room and transported to a separate lab for amplification. The two reactions were purified and eluted in 25  $\mu$ l of TET each over MinElute columns (Qiagen), then assessed for efficiency using a real-time qPCR assay targeting the IS5 and IS6 sequences in the indexing primers. The reactions were then pooled into one double-indexed library. Approximately one third of the library was amplified over three 70  $\mu$ l PCR reactions using 5  $\mu$ l of template each and Herculanase II Fusion DNA Polymerase (Agilent). The products were MinElute purified, pooled, and quantified using an Agilent Tape Station D1000 Screen Tape kit. LUND1 and the corresponding negative controls were sequenced separately on an Illumina NextSeq 500 using single-end, 75-cycle, high-output kits.

#### **Pathogen identification and authentication**

De-multiplexed sequencing reads belonging to LUND1 were processed in silico with the EAGER pipeline (v.1.92) [24]. ClipAndMerge was used for adapter removal, fragment length filtering (minimum sequence length, 30 bp), and base sequence quality filtering (minimum base quality, 20). MALT v. 038 [19] was used to screen the metagenomic data for pathogens using the full NCBI Nucleotide database (“nt,” April 2016) with a minimum percent identity of 85%, a minSupport threshold of 0.01, and a topPercent value of 1.0. The resulting metagenomic profile was visually assessed with MEGAN6 CE [20]. The adapter-clipped reads were additionally aligned to a reconstructed MTBC ancestor genome [21] with BWA [33] as implemented in EAGER (-l 1000, -n 0.01, -q 30). Damage was characterized with DamageProfiler in EAGER [84].

#### **In-solution capture probe design**

Single-stranded probes for in-solution capture were designed using a computationally extrapolated ancestral genome of the MTBC [21]. The probes are 52 nucleotides in length with a tiling density of 5 nucleotides, yielding a set of 852,164 unique probes after the removal of duplicate and low complexity probes. The number of probes was raised to 980,000 by a random sampling among the generated probe sequences. A linker sequence (5'-CACTGCGG-3') was attached to each probe sequence, resulting in probes of 60 nucleotides in length, which were printed on a custom-design 1 million-feature array (Agilent). The printed probes were cleaved off the array, biotinylated, and prepared for capture according to Fu et al. [30].

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

Fifty microliters of the original LUND1 extract were used to create a uracil-DNA glycosylase (UDG) treated library, in which the post-mortem cytosine to uracil modifications, which cause characteristic damage patterns in ancient DNA, are removed. The template DNA was treated in a buffer including 7  $\mu$ l H<sub>2</sub>O, 10  $\mu$ l NEB Buffer 2 (New England Biolabs), 12  $\mu$ l dNTP mix (2.5 mM), 1  $\mu$ l BSA (10 mg/ml), 10  $\mu$ l ATP (10 mM), 4  $\mu$ l T4 polynucleotide kinase, and 6  $\mu$ l USER enzyme (New England Biolabs). The reaction was incubated at 37 °C for 3 h, and then 4  $\mu$ l of T4 polymerase was added to the

library to complete the blunt-end repair step. The remainder of the library preparation protocol, including double indexing, was performed as described above.

The LUND1 UDG-treated library was amplified over two rounds of amplification using Herculase II Fusion DNA Polymerase (Agilent). In the first round, five reactions using 3  $\mu$ l of template each were MinElute purified and pooled together. The second round of amplification consisted of three reactions using 3  $\mu$ l of template each from the first amplification pool. The resulting products were MinElute purified and pooled together. The final concentration of 279 ng/ $\mu$ l was measured using an Agilent Tape Station D1000 Screen Tape kit (Agilent). A portion of the non-UDG library (see above) was re-amplified to 215 ng/ $\mu$ l. A 1:10 pool of the non-UDG and UDG amplification products was made to undergo capture. A pool of all associated negative control libraries (Supplementary Table 2) and a positive control known to contain *M. tuberculosis* DNA also underwent capture in parallel with the LUND1 libraries. Capture was performed according to an established protocol [29], and the sample product was sequenced on an Illumina NextSeq with a 150-cycle paired end kit to a depth of ~ 60 million paired reads. The negative controls were sequenced on a NextSeq 500 with a 75-cycle paired end kit.

#### Genomic reconstruction, heterozygosity, and SNP calling

For the enriched, UDG-treated LUND1 sequencing data, de-multiplexed paired-end reads were processed with the EAGER pipeline (v. 1.92) [24], adapter-clipped with AdapterRemoval, and aligned to the MTBC reconstructed ancestor genome with BWA (-l 32, -n 0.1, -q 37). Previously published ancient and modern *Mycobacterium tuberculosis* genomic data (Supplementary Table 4, Supplementary Table 5) were processed as single-end sequencing reads, but otherwise processed identically in the EAGER pipeline. Genome Analysis Toolkit (GATK) UnifiedGenotyper was used to call SNPs using default parameters and the EMIT ALL SITES output option [85]. We used MultiVCFAnalyzer (v0.87 <https://github.com/alexherbig/MultiVCFAnalyzer>) [5] to create and curate SNP alignments for the L4 (Supplementary Table 5) and full MTBC (Supplementary Table 4) datasets based on SNPs called in reference to the TB ancestor genome [21], with repetitive sequences, regions subject to cross-species mapping, and potentially imported sites excluded. The repetitive and possibly cross-mapped regions were excluded as described previously [5]. Potentially imported sites were identified using ClonalFrameML [34] separately for each dataset, using full genomic alignments and trees generated in RAxML [86] as input without the respective outgroups. Remaining variants were called as homozygous if they were covered by at least 5 reads, had a minimum genotyping quality of 30, and constituted at least 90% of the alleles present at the site. Outgroups for each dataset were included in the SNP alignments, but no variants unique to the selected outgroup genomes were included. Minority alleles constituting over 10% were called and assessed for LUND1 to check for a multiple strain *M. tuberculosis* infection. Sites with missing or incomplete data were excluded from further analysis.

#### Phylogenetic analysis

Neighbor joining (Figs. S2 and S3 in Additional File 3), maximum likelihood (Figs. S4 and S5 in Additional File 3), and maximum parsimony (Figs. S6 and S7 in Additional

File 3) trees were generated for the L4 and full MTBC datasets (Tables S4 and S5 in Additional File 1), with 500 bootstrap replications per tree. Maximum parsimony and neighbor joining trees were configured using MEGA-Proto and executed using MEGA-CC [66]. Maximum likelihood trees were configured and executed using RAxML [86] with the GTR+GAMMA (4 gamma categories) substitution model.

#### Bayesian phylogenetic analysis of full MTBC and L4 datasets

Bayesian phylogenetic analysis of the full MTBC was conducted using a dataset of 261 *M. tuberculosis* genomes including LUND1, five previously published ancient genomes [5, 6], and 255 previously published modern genomes (Additional File 5). *Mycobacterium canettii* was used as an outgroup for this dataset. Bayesian phylogenetic analysis of L4 of the MTBC was conducted using a dataset of 152 genomes including three ancient genomes presented here and in a previous publication [6] and 149 previously published modern genomes (Additional File 6). Body80 and body92 were selected out of the eight samples presented by Kay and colleagues based on multiple criteria. Multiple samples from that study proved to be mixed strain infections. Apart from body92, these samples were excluded from this analysis due to our present inability to separate strains with retention of unique positions. Body92 had a clearly dominant strain estimated by Kay et al. [6] to constitute 96% of the tuberculosis DNA in the sample, and stringent mapping in BWA [33] (-l 32, -n 0.1, -q 37) for this project found the genome to have 124-fold coverage when mapped against the TB ancestor. Between the degree of dominance and the high coverage, we could confidently call variant positions from the dominant strain (Fig. S16a in Additional File 3). Body80 was the only single-strain sample from that collection to have sufficient coverage (~8x) for confident SNP calling after stringent mapping (Fig. S16b in Additional File 3). For selection criteria for the modern genomes, please see Additional File 3. L2\_N0020 was used as an outgroup. The possibility of equal evolutionary rates in both datasets was rejected by the MEGA-CC molecular clock test [66]. TempEst [87] was also used to assess temporal structure in the phylogeny prior to analysis with BEAST2 [37]. For the full MTBC alignment,  $R^2 = 0.273$ , and for the L4 alignment,  $R^2 = 0.113$  (Figs. S8 and S9 in Additional File 3). We generated a maximum likelihood tree and alignment for the full MTBC excluding the animal-associated lineages, and consequently excluding the ancient *M. pinnipedii* genomes, to test if limiting the dataset to the human lineages produced a stronger temporal signal. Without the anchor of the ancient *M. pinnipedii* genomes, the temporal signal for the full complex reduced ( $R^2 = 0.06$ ), as all ancient calibration points were limited to Lineage 4 (Fig. S17 in Additional File 3). When the root-to-tip distances are plotted with points labeled according to lineage or sublineage, it becomes clear that clade membership is largely driving the distance from root of the genomes. However, there remains a temporal signal in the data.

A correction for static positions in the *M. tuberculosis* genome not included in the SNP alignment was included in the configuration file. A “TVM” substitution model, selected based on results from ModelGenerator [88], was implemented in BEAUti as a GTR+G4 model with the AG rate parameter fixed to 1.0. LUND1, body80, and body92 were tip-calibrated using year of death, which was available for all three individuals (Additional File 6). The three ancient Peruvian genomes were calibrated using the midpoint of their OxCal ranges (Additional File 5) [5]. We performed tip sampling for all

modern genomes excluding the outgroup over a uniform distribution between 1992 and 2010 in all models. The outgroup was fixed to 2010 in every case. All tree priors were used in conjunction with an uncorrelated relaxed lognormal clock model. The constant coalescent model was also used in conjunction with a strict clock model.

Two independent MCMC chains of 200,000,000 iterations minimum were computed for each model. If the ESS for any parameter was below 200 after the chains were combined, they were resumed with additional iterations. The results were assessed in Tracer v1.7.1 with a 10% burn-in [89]. Trees were sampled every 20,000 iterations. The log files and trees for each pair of runs were combined using LogCombiner v2.4.7 [37]. An MCC tree was generated using TreeAnnotator with 10% burn-in [37]. Figures 3 and 4 were generated using the ggtree package [90] in R [91]. For details on the parameterization of the birth-death models, please see Additional File 3.

Marginal likelihood was calculated using stepping stone sampling [92] implemented in the MODELSELECTION package in BEAST2. The total chain length required for convergence in each model was split across 100 steps. Following this, we performed a date randomization test [41] for the BDSKY+UCLD model for each dataset. Dates were shuffled randomly among all genomes excluding the outgroup. For both datasets the outgroup was used as an anchor for tip-dating of the “modern” genomes in each date-randomized model. Ten randomizations were generated for each model and run in at least two parallel chains. For the L4 dataset, the chains were run until the rate parameter reached an ESS of at least 200 for every date-randomized model (Additional File 11). For the date randomizations of the full MTBC dataset, we reached sufficient ESS in four out of ten models. However, as noted above in the “Results” section, we reached ESS values greater than or equal to 100 for the rate parameter for all models. We present the rate estimates and rate parameter ESS values for all MTBC date randomizations (Additional File 10).

## Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13059-020-02112-1>.

**Additional file 1: Table S1.** Assigned reads from all taxonomic levels represented in the metagenomic LUND1 library prior to in-solution capture for MTBC DNA.

**Additional file 2: Table S2.** Full EAGER pipeline results for LUND1 shotgun sequencing data when mapped to HG19 human reference genome and TB ancestor genome, the non-UDG-treated enriched LUND1 data when mapped to the TB ancestor genome, and the UDG-treated enriched LUND1 data when mapped to the TB ancestor genome.

**Additional file 3: Supplementary information.** Detailed supplements to the RESULTS and METHODS sections, including supplementary figures.

**Additional file 4: Table S3.** Full EAGER pipeline results for negative controls processed with LUND1, mapped to the reconstructed TB ancestor genome.

**Additional file 5: Table S4.** List of genomes included in the full MTBC dataset, with respective publications, accession numbers, lineages, and dates (when applicable).

**Additional file 6: Table S5.** List of genomes included in the L4 dataset, with respective publications, accession numbers, lineages, dates, and percentage of total SNPs called as heterozygous.

**Additional file 7: Table S6.** List of sites excluded from the full MTBC dataset in GFF format.

**Additional file 8: Table S7.** List of sites excluded from the L4 dataset in GFF format.

**Additional file 9: Table S8.** SnpEff annotation for derived alleles in LUND1.

**Additional file 10: Table S9.** Date randomization results for MTBC dataset.

**Additional file 11: Table S10.** Date randomization results for L4 dataset.

**Additional file 12.** Review history.

### Acknowledgements

The authors would like to thank Marta Burri for conducting the hybridization capture. We also thank Elizabeth A. Nelson for laboratory assistance; Maria Sprou, Felix M. Key, and Luis Roger Esquivel Gomez for technical assistance; and all members of the Molecular Palaeopathology and Computational Pathogenomics research groups at the MPI-SHH for constructive comments throughout the study. We would like to acknowledge the Department of Imaging and Clinical Physiology, Skåne University Hospital Lund, for the generous opportunity to examine the mummy of Peder Winstrup and the radiologists Roger Siemund, Pär Wingren, Mats Geijer, Pernilla Gustavson and David Pellby for contributing to the image analyses. A sincere thank you to the thoracic surgeons Erik Gyllstedt and Jesper Andreasson for so delicately extracting the small calcification of interest and enabling further investigations.

### Review history

The review history is available as Additional file 12.

### Peer review information

Andrew Cosgrove was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### Authors' contributions

C.A. and K.I.B. conceived of the investigation. S.S., D.K., A.H., and K.I.B. designed the experiments. T.A., G.B., and C.A. performed the exhumation and radiological analysis of the mummy and provided a paleopathological examination. G.B. was responsible for the CT examinations together with imaging analysis and coordination of the calcification extraction. S.S. performed laboratory work. S.S., D.K., A.H., Å.J.V., and K.I.B. performed analyses. All authors have read and approved this manuscript.

### Funding

This project was supported by the Max Planck Society and by grants from the Erik Philip-Sörensen Foundation and Crafoord foundation to T.A. and C.A. Open access funding provided by Projekt DEAL.

### Availability of data and materials

Raw sequencing data generated within this study was uploaded to the NCBI Sequence Read Archive (SRA) (accession SRS6462469; BioProject PRJNA517266) [93]. These data include the non-UDG non-enriched screening library, the non-UDG enriched library, and the UDG-treated enriched library. The previously published data used in the MTBC dataset is available on the SRA and can be accessed as part of the following BioProject accessions: PRJNA244165 [94], PRJEB7454 [95], PRJNA186722 [96], PRJEB3128 [97], PRJEB3223 [98], PRJNA52007 [97], PRJNA39969 [100], PRJEB2138 [101], PRJNA52637 [102], PRJNA38491 [103], PRJNA49659 [104], PRJEB2092 [105], PRJEB2091 [106], and PRJNA244633 [107]. See Additional File 5 for sample-specific accession numbers. The previously published data used in the L4 dataset is available on the SRA and can be accessed as part of the following BioProject accessions: PRJEB7454 [95], PRJEB11460 [108], PRJEB3223 [98], PRJNA52007 [99], PRJNA52637 [102], PRJNA38491 [103], PRJNA39969 [100], and PRJNA49659 [104]. See Additional File 6 for sample-specific accession numbers.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

Not applicable.

### Author details

<sup>1</sup>Department of Archaeogenetics, Max Planck Institute for the Science of Human History, 07745 Jena, Germany.

<sup>2</sup>Present address: Section for Evolutionary Genomics, The GLOBE Institute, University of Copenhagen, 1353

Copenhagen, Denmark. <sup>3</sup>Department of Archaeology and Ancient History, Lund University, 221 00 Lund, Sweden.

<sup>4</sup>Department of Medical Imaging and Clinical Physiology, Skåne University Hospital Lund and Lund University, 221 00

Lund, Sweden. <sup>5</sup>Arkeologerna, National Historical Museum, 226 60 Lund, Sweden. <sup>6</sup>Transmission, Infection,

Diversification & Evolution Group, Max Planck Institute for the Science of Human History, 07745 Jena, Germany.

Received: 9 August 2019 Accepted: 17 July 2020

Published online: 10 August 2020

### References

1. WHO. WHO | Tuberculosis (TB). WHO. 2018 [cited 2018 Nov 18]. Available from: <http://www.who.int/tb/en/>.
2. Houben RMGJ, Dodd PJ. The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. *PLoS Med*. 2016;13(10):e1002152.
3. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet*. 2013;45(10):1176–82.
4. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, et al. The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLOS Pathogens*. 2013;9(8):e1003543.
5. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*. 2014;514(7523):494–7.



6. Kay GL, Sergeant MJ, Zhou Z, Chan JZ-M, Millard A, Quick J, et al. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun*. 2015;6:6717.
7. Baker O, Lee OY-C, Wu HHT, Besra GS, Minnikin DE, Llewellyn G, et al. Human tuberculosis predates domestication in ancient Syria. *Tuberculosis*. 2015;95:54–12.
8. Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OY-C, Gernaey AM, et al. Detection and molecular characterization of 9000-year-old *Mycobacterium tuberculosis* from a Neolithic settlement in the Eastern Mediterranean. Ahmed N, editor. *PLoS ONE*. 2008;3(10):e3426.
9. Masson M, Molnár E, Donoghue HD, Besra GS, Minnikin DE, Wu HHT, et al. Osteological and biomolecular evidence of a 7000-year-old case of hypertrophic pulmonary osteopathy secondary to tuberculosis from Neolithic Hungary. *PLoS One*. 2013;8(10):e78252.
10. Rothschild BM, Martin L, Lev G, Bercovier H, Bar-Gal GK, Greenblatt CL, et al. *Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present. *Clin Infect Dis*. 2001;33:305–11.
11. Wilbur AK, Bouwman AS, Stone AC, Roberts CA, Pfister L-A, Buikstra JE, et al. Deficiencies and challenges in the study of ancient tuberculosis DNA. *J Archaeol Sci*. 2009;36(9):1990–7.
12. Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol*. 2018;16(4):202–13.
13. Ochman H, Elwyn S, Moran NA. Calibrating bacterial evolution. *PNAS*. 1999;96(22):12638–43.
14. Warinner C, Herbig A, Mann A, Yates JAF, Weiß CL, Burbano HA, et al. A robust framework for microbial archaeology. *Annual Review of Genomics and Human Genetics*. 2017;18(1):null.
15. Leung AN, Müller NL, Pineda PR, FitzGerald JM. Primary tuberculosis in childhood: radiographic manifestations. *Radiology*. 1992;182(1):87–91.
16. Burrill J, Williams CJ, Bain G, Conder G, Hine AL, Misra RR. Tuberculosis: a radiologic review. *Radiographics*. 2007;27(5):1255–73.
17. Bahr NC, Antinori S, Wheat LJ, Sarosi GA. Histoplasmosis infections worldwide: thinking outside of the Ohio River valley. *Curr Trop Med Rep*. 2015;2(2):70–80.
18. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc Natl Acad Sci*. 2013;110(39):15758–63.
19. Vågene ÅJ, Herbig A, Campana MG, García NMR, Warinner C, Sabin S, et al. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature Ecol Evol*. 2018;2:520–8.
20. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol*. 2016;12(6):e1004957.
21. Comas I, Chakravarti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet*. 2010;42(6):498–503.
22. Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, Arlat M, et al. Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature*. 2002;415(6871):497–502.
23. Mann AE, Sabin S, Ziesemer K, Vågene ÅJ, Schroeder H, Ozga AT, et al. Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Sci Rep*. 2018;8(1):9822.
24. Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, et al. EAGER: efficient ancient genome reconstruction. *Genome Biol*. 2016;17:60.
25. Renaud G, Slon V, Duggan AT, Kelso J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol*. 2015;16(1):1–18.
26. Dabney J, Meyer M, Pääbo S. Ancient DNA damage. *Cold Spring Harb Perspect Biol*. 2013;5(7):a012567.
27. Ginothac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*. 2011;27(15):2153–5.
28. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res*. 2010;38(6):e87.
29. Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, Brizuela L, et al. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protocols*. 2009;4(6):960–74.
30. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, et al. DNA analysis of an early modern human from Tianyuan Cave, China. *PNAS*. 2013;110(6):2223–7.
31. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet*. 2016;48:1535–43.
32. Coll F, McNERney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun*. 2014;5:4812.
33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60.
34. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol*. 2015;11(2):e1004041.
35. Boritsch EC, Khanna V, Pawlik A, Honoré N, Navas VH, Ma L, et al. Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria. *PNAS*. 2016;15:201604921.
36. Mortimer TD, Pepperell CS. Genomic signatures of distributive conjugal transfer among mycobacteria. *Genome Biol Evol*. 2014;6(9):2489–500.
37. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 2014;10(4):e1003537.
38. Kühnert D, Coscolla M, Brites D, Stucki D, Metcalfe J, Fenner L, et al. Tuberculosis outbreak investigation using phylodynamic analysis. *Epidemics*. 2018;25:47–53.
39. O'Neill MB, Shockey AC, Zarley A, Aylward W, Eldholm V, Kitchen A, et al. Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia bioRxiv 2018 6:210161.
40. Duchêne S, Duchêne D, Holmes EC, Ho SYW. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol Biol Evol*. 2015;32(7):1895–906.

41. Ramsden C, Holmes EC, Charleston MA. Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Mol Biol Evol.* 2009;26(1):143–53.
42. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms. *SnPEff Fly.* 2012;6(2):80–92.
43. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, et al. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog.* 2008;4(9):e1000160.
44. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* 2008;6(12):e311.
45. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omais B, Marmiesse M, et al. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Pathog.* 2005;1(1):e5.
46. Shillington K. *History of Africa*. Third. New York: Palgrave MacMillan; 2012.
47. Cohen MN, Armelagos GJ. *Paleopathology at the origins of agriculture*. Gainesville: University Press of Florida; 1984.
48. Armelagos GJ, Brown PJ, Turner B. Evolutionary, historical and political economic perspectives on health and disease. *Soc Sci Med.* 2005;61(4):755–65.
49. Roberts CA, Buikstra JE. *The bioarchaeology of tuberculosis: a global view on a reemerging disease*. Gainesville: University Press of Florida; 2003.
50. Canci A, Minozzi S, Tarli SMB. New evidence of Tuberculous spondylitis from Neolithic Liguria (Italy). *Int J Osteoarchaeol.* 1996;6(5):497–501.
51. El-Najjar M, Al-Shiyab A, Al-Sarie I. Cases of tuberculosis at 'Ain Ghazal, Jordan. *Paléorient.* 1996;22(2):123–8.
52. Formicola V, Milanese Q, Scarsini C. Evidence of spinal tuberculosis at the beginning of the fourth millennium BC from Arene Candide cave (Liguria, Italy). *Am J Phys Anthropol.* 1987;72(1):1–6.
53. Köhler K, Pálfi G, Molnár E, Zalai-Gaál I, Osztás A, Bánffy E, et al. A late Neolithic case of Pott's disease from Hungary. *Int J Osteoarchaeol.* 2014;24(6):697–703.
54. Sparacello VS, Roberts CA, Kerudin A, Müller R. A 6500-year-old Middle Neolithic child from Pollera Cave (Liguria, Italy) with probable multifocal osteoarticular tuberculosis. *International Journal of Paleopathology.* 2017 Feb [cited 2017 Feb 13]; Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1879981716300900>.
55. Ho SYW, Phillips MJ, Cooper A, Drummond AJ. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol.* 2005;22(7):1561–8.
56. Ho SYW, Larson G. Molecular clocks: when times are a-changin'. *Trends Genet.* 2006;22(2):79–83.
57. Ho SYW, Shapiro B, Phillips MJ, Cooper A, Drummond AJ. Evidence for time dependency of molecular rate estimates. *Syst Biol.* 2007;56(3):515–22.
58. Achtman M. How old are bacterial pathogens? *Proc R Soc B.* 2016;283(1836):20160990.
59. Duchêne S, Holmes EC, Ho SYW. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc R Soc B.* 2014;281(1786):20140732.
60. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. *Trends Ecol Evol.* 2003;18(9):481–8.
61. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 2006;4(5):e88.
62. Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, et al. Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Science Advances.* 2018;4(10):eaat5869.
63. O'Neill MB, Shockey A, Zarley A, Aylward W, Eldholm V, Kitchen A, et al. Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. *Mol Ecol.* 2019;mec.15120.
64. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. *PNAS.* 2006;103(8):2869–73.
65. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, et al. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet.* 2011;43(5):482–6.
66. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics.* 2012;28(20):2685–6.
67. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *PNAS.* 2013;110(1):228–33.
68. Boskova V, Stadler T, Magnus C. The influence of phylodynamic model specifications on parameter estimates of the Zika virus epidemic. *Virus Evol.* 2018 1 [cited 2018 Nov 2];4(1). Available from: <https://academic.oup.com/ve/article/4/1/vex044/4829709>.
69. Möller S, Plessis L du, Stadler T. Impact of the tree prior on estimating clock rates during epidemic outbreaks. *PNAS* 2018 28; 201713314.
70. Brown K, Mund DF, Aberle DR, Batra P, Young DA. Intrathoracic calcifications: radiographic features and differential diagnoses. *RadioGraphics.* 1994;14(6):1247–61.
71. Kay GL, Sergeant MJ, Giuffra V, Bandiera P, Milanese M, Bramanti B, et al. Recovery of a Medieval *Brucella melitensis* Genome Using Shotgun Metagenomics. *mBio.* 2014;5(4):e01337–e013314. <https://doi.org/10.1128/mBio.01337-14>.
72. Devault AM, Mortimer TD, Kitchen A, Kieseewetter H, Enk JM, Golding GB, et al. A molecular portrait of maternal sepsis from Byzantine Troy. *eLife.* 2017;6. <https://doi.org/10.7554/eLife.20983>.
73. Jankute M, Nataraj V, Lee OY-C, Wu HHT, Ridell M, Garton NJ, et al. The role of hydrophobicity in tuberculosis evolution and pathogenicity. *Sci Rep.* 2017;7(1):1315.
74. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, et al. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *PNAS.* 1997;94(18):9869–74.
75. Dubos R, Dubos J. *The white plague: tuberculosis, man, and society*. 3rd ed. New Brunswick: Rutgers University Press; 1952.
76. Gansauge M-T, Gerber T, Glocke I, Korlević P, Lippik L, Nagel S, et al. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 2017;45(10):e79.
77. Andrades Valtueña A, Mitnik A, Key FM, Haak W, Allmäe R, Belinskij A, et al. The Stone Age plague and its persistence in Eurasia. *Current Biology.* 2017 22 [cited 2017 Nov 30]; Available from: <http://www.sciencedirect.com/science/article/pii/S0960982217313283>.

78. Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren K-G, et al. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell*. 2015;163(3):571–82.
79. Spyrou MA, Tukhbatova RI, Wang C-C, Valtueña AA, Lankapalli AK, Kondrashin VV, et al. Analysis of 3800-year-old *Yersinia pestis* genomes suggests Bronze Age origin for bubonic plague. *Nat Commun*. 2018;9(1):2234.
80. van de Loosdrecht M, Bouzouggar A, Humphrey L, Posth C, Barton N, Aximu-Petri A, et al. Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. *Science*. 2018;360(6388):548–52.
81. Schuenemann VJ, Peltzer A, Welte B, van Pelt WP, Molak M, Wang C-C, et al. Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nat Commun*. 2017;8:15694.
82. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*. 2010;2010(6):pdb.prot5448-pdb.prot5448.
83. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res*. 2012;40(1):e3.
84. Neukamm J, Peltzer A. Integrative-Transcriptomics/DamageProfiler. 2018. Available from: <https://doi.org/10.5281/zenodo.1291355>.
85. der Auwera GAV, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43(1):11.10.1–11.10.33.
86. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312–3.
87. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*. 2016 [cited 2017 Nov 30];2(1). Available from: <https://academic.oup.com/ve/article/2/1/vew007/1753488>.
88. Keane T, Creevey C, Pentony M, Naughton TJ, McInerney J. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol*. 2006;6:29–46.
89. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*. 2018;67(5):901–4. <https://doi.org/10.1093/sysbio/syy032>.
90. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. 2017;8(1):28–36.
91. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. Available from: <https://www.R-project.org/>.
92. Xie W, Lewis PO, Fan Y, Kuo L, Chen M-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst Biol*. 2011;60(2):150–60.
93. Sabin S, Herbig A, Vågene AJ, Ahlström T, Bozovic G, Arcini C, Kühnert D, Bos KI. A seventeenth-century *Mycobacterium tuberculosis* genome supports a Neolithic emergence of the *Mycobacterium tuberculosis* complex. *Datasets*. NCBI Bioproject. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA517266> (2020).
94. University of Tuebingen. *Mycobacterium tuberculosis* genome sequencing. National Library of Medicine (US), National Center for Biotechnology Information. 2014. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA244165>.
95. Warwick University. Metagenome analysis of TB in mummies, molecular analysis of TB in 18th century mummy samples from Vac, Hungary. National Library of Medicine (US), National Center for Biotechnology Information. 2014. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJEB7454>. Accessed in 2016.
96. Ludwig-Maximilians-University. *Mycobacterium tuberculosis* variant caprae Genome sequencing. National Library of Medicine (US), National Center for Biotechnology Information. 2013. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA186722>. Accessed in 2014.
97. STH. *Mycobacterium tuberculosis* complex genetic diversity. National Library of Medicine (US), National Center for Biotechnology Information. 2013. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJEB3128>. Accessed in 2014.
98. CSISP. Global diversity of *Mycobacterium tuberculosis* complex isolates - MTBC reference dataset - Single end data. National Library of Medicine (US), National Center for Biotechnology Information. 2013. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJEB3223>. Accessed in 2014.
99. Broad Institute. *Mycobacterium tuberculosis* strain: WGS\_comparative, Multi-isolate study of *Mycobacterium tuberculosis*. National Library of Medicine (US), National Center for Biotechnology Information. 2010. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA52007>. Accessed in 2014.
100. Broad Institute. *Mycobacterium* strain: Phylogenetic\_Comparative, Comparative Genomics of *Mycobacterium tuberculosis* and *Mycobacterium canetti*. National Library of Medicine (US), National Center for Biotechnology Information. 2009. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA39969>. Accessed in 2014.
101. SC. Discovery of sequence diversity in *Mycobacterium tuberculosis* (Russia collection). National Library of Medicine (US), National Center for Biotechnology Information. 2010. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJEB2138>. Accessed in 2014.
102. Harvard School of Public Health, Department of Immunology and Infectious Diseases. *Mycobacterium tuberculosis* strain: MicroEvolution, MicroEvolution study of *Mycobacterium tuberculosis*. National Library of Medicine (US), National Center for Biotechnology Information. 2010. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA52637>. Accessed in 2014.
103. Broad Institute. *Mycobacterium tuberculosis* str. Erdman = ATCC 35801, *Mycobacterium tuberculosis* Erdman genome sequencing. National Library of Medicine (US), National Center for Biotechnology Information. 2009. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA38491>. Accessed in 2014.
104. Simon Fraser University. *Mycobacterium tuberculosis* BC strain: BC, *Mycobacterium tuberculosis* BC genome sequencing. National Library of Medicine (US), National Center for Biotechnology Information. 2010. Available from: [www.ncbi.nlm.nih.gov/bioproject/PRJNA49659](http://www.ncbi.nlm.nih.gov/bioproject/PRJNA49659). Accessed in 2014.
105. SC. The genome sequence of antelope associated clade of the *Mycobacterium tuberculosis* complex (strain VIC4031ORYX). National Library of Medicine (US), National Center for Biotechnology Information. 2010. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJEB2092>. Accessed in 2014.

106. SC. Mycobacterium microti genome diversity project. National Library of Medicine (US), National Center for Biotechnology Information. 2011. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJEB2091>. Accessed in 2014.
107. University of Tuebingen. *Mycobacterium tuberculosis* variant pinnipedii, Mycobacterium pinnipedii Genome sequencing. National Library of Medicine (US), National Center for Biotechnology Information. 2014. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA244633>. Accessed in 2017.
108. University of Basel. MTB sublineage evolution, Evolution of *Mycobacterium tuberculosis* sublineages. National Library of Medicine (US), National Center for Biotechnology Information. 2016. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJEB11460>.

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

