Connection between two historical tuberculosis outbreak sites in Japan, Honshu, by a new ancestral *Mycobacterium tuberculosis* L2 sublineage

Christophe Guyeux¹, Gaetan Senelle¹, Guislaine Refrégier²,³, Florence Bretelle-Establet⁴, Emmanuelle Cambau⁵,⁶, Christophe Sola²,⁵*,

¹FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department, Univ. Bourgogne Franche-Comté (UBFC), 16 Route de Gray, 25000 Besançon
²Université Paris-Saclay, Saint-Aubin, France
³Ecologie Systematique Evolution, Batiment 360, Université Paris-Saclay, CNRS, AgroParisTech, Orsay 91400, France
⁴Université de Paris, SPHERE, UMR7219, CNRS, Paris, France
⁵Université de Paris, IAME, UMR1137, INSERM, Paris, France
⁶AP-HP, GHU Nord, service de mycobactériologie spécialisée et de référence, Laboratoire associé du Centre National de Référence des mycobactéries et résistance des mycobactéries aux antituberculeux (CNR-MyRMA), Paris, France

This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.
Summary

By gathering 680 publicly available Sequence Read Archives from isolates of *Mycobacterium tuberculosis* complex (MTBC) including 190 belonging to the lineage 2 *Beijing*, and using an in-house bioinformatical pipeline, the *TB-Annotator*, that analyzes more than 50,000 characters, we describe herein a new L2 sublineage from 20 isolates found in the Tochigi province, (Japan), that we designate as *asia ancestral 5* (AAnc5). These isolates harbor a number of specific criteria (42 SNPs) and their intra-cluster pairwise distance suggests historical and not epidemiological transmission. These isolates harbor a mutation in *rpoC*, and do not fulfill, any of the *modern Beijing* lineage criteria, nor any of the other *ancestral Beijing* lineages described so far. *Asia ancestral 5* isolates do not possess *mutT2* 58 and *ogt* 12 characteristics of *modern Beijing*, but possess *ancestral Beijing* SNPs characteristics. By looking into the literature, we found a reference isolate ID381, described in Kobe and Osaka belonging to the “G3” group, sharing 36 out of the 42 specific SNPs found in AAnc5. We also assessed the intermediate position of the *asia ancestral 4* (AAnc4) sublineage recently described in Thailand and propose an improved classification of the L2 that now includes AAnc4 and AAnc5. By increasing the recruitment into *TB-Annotator* to around 3000 genomes (including 642 belonging to L2), we confirmed our results and discovered additional historical *ancestral* L2 branches that remain to be investigated in more detail. We also present, in addition, some anthropological and historical data from Chinese and Japan history of tuberculosis, as well as from Korea, that could support our results on L2 evolution. This study shows that the reconstruction of the early history of tuberculosis in Asia is likely to reveal complex patterns since its emergence.

Introduction

With 9.9 million new cases in 2019, and 500,000 multi-drug resistant cases, tuberculosis (TB) is far from being eradicated [1]. Among 9 acknowledged lineages (L1 to L9) described in *Mycobacterium tuberculosis* complex, the lineage L2 is of great interest [2-6]. Very large outbreaks in L2 were shown to have independently emerged worldwide [7]. Although L2 origin is suspected to be in China and L2 is predominant in east Asia, its exact place and time of emergence are still highly debated [8-10]. L2 clinical isolates have developed specific virulence and drug-resistance features that contribute to their epidemic
success [6, 11-13]. From an evolutionary standpoint, L2 has developed a high IS6110 copy number lifestyle, that could have fostered an hypermutator phenotype that may have increased the virulence or fitness of some of these isolates [14, 15]. The epidemic burst of L2 was first detected during the 90s and was promoted by historical and geopolitical events such as: (1) the fall of the former USSR and changes in China; (2) a poorly individualized TB treatment and medical follow-up of prisoners in these countries; (3) the increase in the global trade share of China [16-18]. L2 fostered many research studies to understand multidrug resistance emergence and to improve its genetic characterization [4, 19, 20]. This was achieved progressively and through the combination of polymorphic markers analysis such as IS6110-RFLP, MIRU-VNTR, Regions of Difference (RD), hypervariable VNTR loci, and lastly whole-genome sequencing (WGS) [21-27].

The L2 lineage was split into two main sublineages, L2.1 (Proto-Beijing) and L2.2 (Beijing), [28-31]. L2.1 was described mainly in south-China, particularly in the Guangxi province and could be as ancient as 30,000 years old, having co-evolved with east Asian populations [8]. Among its characteristics, it harbors the RD105 deletion but not RD207 [32]. Rare L2.1 isolates were shown to have become ultra-drug resistant [32]. L2.2, is defined by SIT1 or variants, and is composed of several sublineages, ancestral and modern ones. L2.2.2 defines the asia ancestral 1 sublineage; L2.2.1 gathers all others [33]. RD181 is specifically deleted in all L2.2.1 sublineages [28]. The switch from ancestral to modern Beijing is associated to the presence of at least one IS6110 copy in the so-called NTF region and the presence of mutations in replication-repair-recombination (3R) genes, among which the mutT2 G->C mutation in position 1286766 and the ogt C->T mutation at codon 12 position 1477596, relatively to the MTBC H37Rv reference sequence (NC_000962.3) [11, 28, 34]. Until recently ancestral Beijing included 3 asia ancestral lineages (asia ancestral 1, 2 and 3, AAnc1, AAnc2, AAnc3) [28], until a new asia ancestral 4 (AAnc4), was discovered in northern Thailand [35]. Modern Beijing strains are responsible of most but not of all of the recent MDR-TB outbreaks [4, 15, 36].

The large databases built with WGS data allow to develop a precise and comprehensive knowledge of L2 diversity [4, 26, 31, 37]. Computational genomics now allows to study in-depth both the global and the local L2 history [6, 24, 26, 32, 35, 38-42]. The evolutionary history of L2 isolates is still debated as is their precise dating of emergence.

---

1 Restriction Fragment Length Polymorphism
2 Mycobacterial Interspersed Repetitive Units-Variable Number of Tandem Repeats
3 Spoligotyping-international-Type
and their geographical origin [26, 34]. Luo et al. suggested that L2 could be as ancient as 30,000 years old [8]. Merker et al. estimated the time to the most recent common ancestor (TMRCA) of ~6,100 and 5,200 years before present (BP) for MIRU-VNTR defined clonal complexes BL7 and CC6 (the most ancient) and ~1,500 years BP for CC5 (the most recent)[26]. Liu et al. estimated the coalescence between L2.1/Proto-Beijing and L2.2.2/ancestral Beijing at 2,200 BP, and at 1,300 years BP for the split between all ancestral Beijing lineages [10]. L2.1/Proto-Beijing expansion would have taken place 900 BP whereas modern Beijing would have appeared only around 500 years ago [10]. Hirsch et al. suggested that east asian and Philippines human populations carrying distinct MTBC lineages may have split only 240–1000 years ago [43].

The geographic origin of the emergence of L2 is as blurry as its dating. According to some authors, north-central and north-east China could have been the initial spreading center [44]. According to others, based on differences in the prevalence of the ancestral L2 lineages in China and on the higher genetic diversity observed in the south-west province of Guizhou, south-China could be the cradle of L2 [8, 34]. Indeed, Guizhou counts 17 ethinical minorities and most of the acknowledged ethnic groups of China are located in this province. More generally, south-east Asia shows a higher human genetic diversity than the north-east Asia [45]. Arguing in favour of a south-China origin, the recent description of an “asia ancestral 4” L2 sublineage, was done in the north of Thailand in Chiangrai, inhabited since the 7th century and peopled by ethnic minorities originating from South-China [35].

Skeletal evidence of tuberculosis during the Bronze age was found in Korea and Japan [46, 47]. In Japan, one of the main characteristics of the tuberculosis history, especially in aged people that were not vaccinated by BCG, is the presence of still poorly characterized ancestral L2 strains [22, 26, 39, 42, 48]. MIRU-VNTR diversity had been shown earlier to be quite important in L2 isolates from Japan [44]. Other evidence based on MIRU-VNTR had suggested that some specific ancestral L2 strains could be endemic in Japan [49-52]. Since the publication of these studies, WGS of a few reference isolates were published in Kobe and Osaka however they were not mentioned in the simplified L2 phylogeny [28, 39].

We are currently developing the TB-Annotator project, a new computational genomics pipeline whose aim is to perform data-mining of MTBC genomic diversity using Sequence Read Archives (SRAs). We studied a set of L2 isolates from central Japan Honshu prefecture, Tochigi [42]. The 169 clinical isolates of M. tuberculosis we studied were from TB patients diagnosed in 2007 and in 2013 [42]. WGS data on these isolates was released after analysis using a bioinformatical pipeline, described as “CAST” by Iwai et al. [53]. Since
a comparative genomic analysis of these isolates with other ancestral L2 had not been performed, we included these genomes in our database. We also included SRAs labelled as *asia ancestral 4* [35]. Using our pipeline, we described the genetic characterization of a new L2 sublineage from Japan, named *asia ancestral 5* (AAnc5), that appears to be exclusive from Japan for the time-being. We also assessed the characteristics of the AAnc4 sublineage described in Thailand and provide an updated global evolutionary framework of the L2 lineage [28].

**Material and Methods**

**Brief description of the TB-Annotator pipeline**

The full version of TB-Annotator is going to be released in another article (Senelle *et al.* in preparation). In brief, regarding the set-up of the platform, the processes are the following: SRAs of interest are selected and kept only provided a certain number of conditions that together reinforce the reliability of the data: they must have read length>75bp, clean reads file must be at least 100 Mo, and CRISPR could be reconstructed using CRISPRbuilder-TB [54]. For each SRA, in addition to reconstructing the CRISPR-Cas region using CRISPRbuilder-TB and apart from collecting NCBI information on the genomes, two scripts successively perform the following tasks: (1) search for SNPs according to reference catalogues (Supplementary Table1) totaling more than 50,000 SNPs, including drug-resistance related SNPs, phylogenetic SNPs as per 26 studies contributing to SNP-based classification ([26, 28, 29, 30, 33, 55-75]; (2) search for additional SNPs in each isolate based on H37Rv reference sequence; (3) look for the presence/absence of H37Rv genes as annotated in mycobrowser (https://mycobrowser.epfl.ch/), (4) look for the presence/absence of deletion regions; (5) identify insertion sites of all known insertion sequences in MTBC. The CRISPR locus is rebuilt semi-automatically using a dedicated and previously published script (43/68/360 Spacers format) with assignment of a Spoligo-International-Type (SIT) tag; the application produces an ordered list of spacers/repeat with variants and IS6110 insertion sequences if present [54, 76, 77].

**Selected genomes.**

We downloaded a set of 680 SRAs; these samples were selected to represent TB genomic diversity (L1 to L9) described so far, including recent landmark papers [29, 30]. The list of these SRAs is shown in Supplementary Table2 (list of 680 SRA including 190 L2
SRA); the database was built to represent all L2 sublineages except for the Pacific RD150 sublineage. From [35], we selected 28 SRAs labeled AAnc4. From [42], we initially included 158 SRAs, however 57 SRAs for which the coverage was either too weak or for which it was impossible to rebuild the spoligotype using CRISPR-builder-TB were discarded [54, 77]

**Bioinformatics and Phylogenetical methods**

Final scripts allow to produce a phylogeny based on the list of studied characters and using RAxML and SplitsTree [78, 79]. All computations were performed on the “Mésocentre de Franche-Comté” supercomputer facilities (141 nodes, 2292 cores, 9,27 To memory, 74,2 CPU Power TFlops, 66,4 GPU TFlops), using adequate commands. Apart from the results shown in the Supplementary Table 3, a final phylogenetic tree is displayed graphically and proprietary python scripts allow interactive queries to be made and results to be shown [79]. The current TB-Annotation version includes 6009 genomes and confirmed our results (Senelle et al. in preparation).

**SNP-based classification of L2 sublineages**

In order to assign the 680 selected SRAs into known lineages and sublineages, we used the reference list of markers defined in Supplementary Table 1. A simplified Venn diagram shows the classification of the 190 L2 SRAs into 125 ancestral and 65 modern isolates (Figure 1). SNP-based classification results on the 190 SRA of the L2 isolates, as produced by TB-Annotation are found in the Supplementary Table 3.

**Methods and dataset used to Define a new asia ancestral 5 sublineage (AAnc5)**

After the assessment of existing known MTBC sublineages (Figure 2A and 2B), the unknown branch (Figure 2C) was investigated in detail: in Silico CRISPR locus reconstruction using CRISPR-builder TB [54, 77], in Silico MIRU-VNTR using CAST [53]; SNPs assessment and intra-cluster distance with TB-Annotation; AAnc5 was shown to originate only from SRAs from Japan isolates (Bioproject PRJDB3875). The bioinformatical pipeline, based on its graphical user interface, allows to select and display new exclusive or shared SNPs and specific genetics characteristics, that were further investigated. Previous genomic markers extracted from Wada et al. 2012 were compared to our results and are found in Supplementary Table 4 [39].

**Results**
1. Discovery of a new L2 Sublineage.

We implemented a representative set of SRAs of L2 including 101 extracted from the 169 samples study from the Tochigi province study, onto the TB-Annotator platform [42]. The classification of the 190 L2 studied genomes is shown in a simplified Venn diagram (Figure 1) and the global phylogenetical tree produced is shown on Figure 2A, 2B, 2C. All samples carried the L2 defining SNP (G497491A) and were found in the same phylogenetic branch (Figure 2A). Based on SNP search, and as shown on Figure 2B and 2C and in more detail in Supplementary Table 3, among the ancestral Beijing (n=125) we found Proto-Beijing (L2.1; n=10), asia ancestral 1 (L2.2.2-AAnc1, n=10), asia-ancestral 2 (AAnc2, n=22), asia ancestral 3 (AAnc3, n=9), asia ancestral 4 (AAnc4, n=28), and an unknown branch (suggested as asia ancestral 5 or AAnc5, n=20); there remained 26 unclassified ancestral Beijing. Figure 1 and Supplementary Table 3.

![Venn diagram showing the classification of the 190 L2 SRA studied; note that AAnc4 are close to modern L2](image)

**Figure 1:** Venn diagram showing the classification of the 190 L2 SRA studied; note that AAnc4 are close to modern L2

Ninety-three isolates (including AAnc4), were part of L2.2.1 and were all harboring a mutT2 G1286766C that is traditionally defining a “modern Beijing” characteristic SNP. Out of these 93, however 65 samples only (excluding AAnc4) were harboring the second modern L2 signature, i.e. the C1477596T SNP in ogt, while the 28 AAnc4 did not harbor this SNP [28,
Hence, the designation of *modern* L2 should be kept for isolates harboring these two SNPs simultaneously and not only the *mutT2* G1286766C polymorphism. The other *modern* Beijing (n=65), are further split into *asian african 1* (n=2), *asian african 3* (n=5), *asian african 2* (AAfr2, n=16), *asian african 2-RD142* (AAfr2-RD142, n=1), *central asian* (n=13), Europe/Russia B0/W148 outbreak (n=2) and there were a remaining 26 *unclassified modern Beijing* isolates, that did not fit to any described modern sublineage definition, and that were not investigated further in this study. No Pacific RD150 isolates was included in this study. We present an improved unified L2 classification scheme that includes the recent discovery of AAnc4, our own AAnc5, as well as the “L2.2.1.2” [30], the “K” strain (a member of AAnc1) [80], and the L2.2.E [82], as shown on Figure 3.

**Figure 2.** Left part (1A); *TB-Annotator* unrooted phylogenetic tree on 680 SRA-derived data. L2 samples are shown in blue. Right part (1B); close-up on the Lineage 2 with most known branches named except in red the new unknown *ancestral* Japan sublineage we designate as *asia ancestral 5*. Central part (1C); focus on the unknown Japan lineage.
Figure 3: Unified Mtb lineage 2 dendrogram representing the current L2 sublineages with some of their defining SNPs or genetic markers. The color code tries to superimpose with each author, the size of the circles is arbitrary; this tree tries to provide a simplified unified evolutionary scheme however does not claim to be representative of the full diversity of L2 (redrawn and adapted from [27]).

The reanalysis of the phylogenetical SNP set described by Shitikov et al. confirmed all the phylogenetical SNPs specific of the ancestral L2 sublineages up to the definition of the modern L2. The mutT2 SNP (G1286766C) is a good phylogenetic marker as it is present only in L2 modern sublineages whereas T1892017C, C4137829T and C4393590G (mutT4) are present in some ancestral sublineages. It is clear from the SNPs results and from the phylogenetical tree shown in Figure 2 that the AAnc4 branch could be called intermediate, as suggested by the team who discovered it, and is neither a bona fide modern L2 sublineage nor a truly ancestral one [35]. Conversely, the 20 isolates that we studied from Japan, further designated as AAnc5, all fulfill the ancestral L2 criteria since they do not possess the expected mutT2 (58) SNP nor any other characteristic of modern Beijing. They also branch before AAnc4. In all cases, the clear-cut split of L2 isolates that either possess none, one, or the two mutT2 (58) and ogt (12) SNPs confirms that these two markers are excellent phylogenetical ones. Based on the results obtained, we thus adapted the unified classification scheme of Shitikov et al. to include some of the new recently described sublineages (Figure
3. Among those, are the “K” strain (part of AAnc1), the “L2.2.1.2”, the L2.2.E [30, 31, 80], and the AAnc5 recently designated as L2.2.A [31].

2. Description of a new asia-ancestral 5 (AAnc5) Japan sublineage

2.1 SNP-based results

We further characterized the Japan asia ancestral 5 sublineage (AAnc5) (n=20) that branches between AAnc1 and AAnc3 (Figure 2B and 2C). These isolates can all be defined by many exclusively shared SNPs. The SNP-based classification produced by the TB-ANNOTATOR pipeline on the 190 L2 genomes can be found in Supplementary Table3. In the future versions of TB-ANNOTATOR, reports will be downloadable from a dedicated website (Senelle et al. in preparation). The list of 42 exclusive SNPs, found in 16 of these 20 isolates is shown in Supplementary Table4. An interesting observation is that all the AAnc5 strains harbor a non-synonymous mutation in 765140 (G>C) in rpoC. Three other genomes (2 in other L2 sublineages and one in a L4 sublineage) also carry this rpoC mutation suggesting independent acquisition.

A pairwise distance matrix between AAnc5 isolates was also computed (Supplementary Table5); the current pipeline systematically computes the intra-branch SNP distance for clusters or selections of interest below 100 SRAs (results not shown). The pairwise distance between the AAnc5 samples shows between a minimum of 166 SNPs (between DRR157280 and DRR157281) and a maximum of 439 SNPs distance (between DRR130203 and DRR034366) inside AAnc5 isolates, thus excluding recent transmission (Supplementary Table5). Assuming a mean 0.3 SNP mutation rate per year per genome, these strains might have diverged approximately between 250 to 600 years from their MRCA. If we accept the suggestion of timing of AAnc4 emergence or expansion around the 7th century in Thailand, (start of Chiangrai), then AAnc5 could have been introduced into Japan earlier, in line with archeological information [35, 46, 47]. A second list of 46 exclusively shared SNPs, shared between the two most distant isolates on a specific subbranch of AAnc5 (DRR034381 and DRR130203, pairwise distance: 417 SNP) is also shown in Supplementary Table5.

2.2 in silico spoligotyping and reconstruction of the CRISPR locus structure of AAnc 5 using TB-ANNOTATOR, IS6110 copy number and insertion sites
The twenty studied AAnc5 isolates showed 6 different spoligotype patterns as reconstructed by CRISPR-builder, which was quite an unexpected result for an L2 sublineage (Supplementary Table6); most of these patterns have been previously described in the SITVIT⁴ database (SIT1, SIT190, SIT269, SIT1364, SIT1674), however one remained undefined as SIT”X”. No SNP variants were found in spacers and repeats, but three isolates exhibited duplications: a duplication of sp65 for DRR034478 and SRR130160, and of sp50 for DRR034476 (Supplementary Table6). The phylogeny that can be derived from reconstruction of the CRISPR-Cas structure confirmed the SNPs results: it reveals sporadic deletions of cas genes, Rv2807c, Rv2808c and Rv2813c in some isolates (Supplementary Table7).

AAnc5 strains were harboring between 14 and 22 IS6110 copies, and two specific copies were found in almost all these L2 isolates and not in other L2 sublineages: one copy was found around position 1724419 in Rv1527c (found in 15 of these isolates) and the second one was found, around position 2041756 (found in 19 of these isolates) (Supplementary Table6).

DRR034455, DRR034471 and DRR034476 were showing the same CRISPR structure, however harbored different missing genes (see next paragraph). Using TB-Annotator, 14 of the AAnc5 isolates were predicted to be drug-sensitive and four were harboring mono-resistant mutations, two were MDRs (Supplementary Table8).

### 2.3 Missing genes

Six TB isolates among the 20 AAnc5, apart from showing classical deletions (RD105, RD207, RD181 and PhiRv1), were harboring specific missing genes: as an example, DRR034363 had Rv1081 to 1084c deleted, DRR034416 was missing Rv1523 to Rv1526c. (Supplementary Table7). These deletions confirm that phylogenetically linked MTB genomes sometimes harbor strain-specific dependent deletions regions due to recombination events.

### 2.4 In Silico VNTR copy number computation using CAST and comparison with other isolates from previous studies

---

⁴ Spoligo-International-Database (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/)
No specific 15+9 VNTR signature could be obtained from *in silico* VNTR typing using CAST for any of the AAnc5 isolates [52, 81]. ETRC, QUB26 and QUB4156 could never be *in silico* predicted. Depending on SRA quality, between 6 to 20 VNTR could be predicted (*Supplementary Table*9). The VNTR results showed slight variation between isolates; eleven VNTR Loci were invariant in this collection (MIRU04, MIRU10, MIRU16, MIRU20, Mtub29, Mtub30, ETRB, MIRU24, MIRU27, Mtub34, MIRU39) whereas nine loci showed variation (MIRU02, MIRU40, Mtub21, QuB11b, ETRA, MIRU23, MIRU26, MIRU31, Mtub39). When comparing with an in-depth VNTR study performed earlier, AAnc5 was shown to belong to M10 or M37 respectively found in Russia and Singapore [44]. When comparing with a set of 5 reference Japanese isolates (A05N056, ID381, 4558, 4994, 4991/M) that were described to represent the main L2 sublineages found in Japan, ID381 was sharing the same VNTR copy number with AAnc5 on 9 loci (*Supplementary Table*9) [22, 39]. When comparing *in silico* VNTR results with previous VNTR results from published studies in Korea, on the “K” strain, known to belong to AAnc1, we retrieved relatively poor similarity [38, 80] (*Supplementary Table*9).

2.5 Comparison between *TB-Annotator* and CAST Server results on prospective Drug Susceptibility Testing and on spoligotyping results.

When comparing the drug-susceptibility testing results obtained using either the *TB-Annotator* pipeline or CAST, they were identical (*Supplementary Table*8 and *Supplementary Table*9). Identical results were also obtained on the classical 43 spacers-format spoligotype reconstruction with a minor and yet unexplained discrepancy on a single spacer of a single isolate, DRR034366, for which CAST predicted SIT250 whereas *TB-Annotator* predicted SIT290 (*Supplementary Table*9).

2.6 AAnc5 is identical to the G3 endemic L2 ancestral strains cluster in Japan

We compared the 42 SNPs table found in the AAnc5 sublineage (*Supplementary Table*4) with the ones found in sequences of 5 Japanese reference isolates, compared to the K1-K2 epidemic strain [39]. According to the definition made by these authors 5 L2 subgroups (G1/2, G3, G4, G5/6 and M) could be defined in Japan based on 10 phylogenetically selected SNPs. From the ID381 strain, a member of the G3 genetic group, described for the first time in Kobe and Osaka in 2006 [39], looking at G3 specific SNPs set
(Supplementary file of [39]), and comparing it with AAnc5 SNPs set, we concluded that the G3 group was new and did not fit with the former Shitikov et al. classification scheme. Indeed, AAnc1, AAnc2 and AAnc3 were found to respectively match with G1/2, G4 and G5/6 in [39], however no equivalent was found for G3.

By comparing SNPs list, we found that the Tochigi province AAnc5 cluster of strains was sharing 36 out of 42 SNPs with the G3:ID381 reference strain found in Kobe and Osaka (Supplementary Table4). 6 SNPs only (C587945T, G765140C, G1202113A, AGGGAG1476812A, G3148446C and G3820365A) were not found in the ID381 strain. We concluded that the Tochigi strains were highly likely to be historically related to the Kobe and Osaka G3 group described in 2006 through the reference isolate ID381. Accordingly, we propose to retain the common SNPs described by Wada et al. and this study as characteristic of the AAnc5 to fit with Shitikov et al. nomenclature. We positioned both AAnc5 and AAnc4 in the global schematic L2 tree [28, 35]. By digging more in-depth into the comparative SNPs list between our study and the former Kobe-Osaka study, we found that DRR034489 was the closest isolate to the G3 ID381 reference sharing 40 more SNP exclusive to the G3 group, whereas another cluster of 3 genomes were more distant but were sharing 15 more SNPs with ID381 (results not shown). As mentioned above, two very distant genomes, DRR034381 and DRR130203 (417 SNPs pairwise distance) where also sharing 46 additional SNPs that were not found elsewhere (Supplementary Table 5).

Discussion

We describe in this paper an historical endemic ancestral sublineage of L2 based on samples collected in central Japan, Tochigi Prefecture, former Shimotsuke province, that we named AAnc5. This sublineage is highly related to the Japanese G3 group defined in 2012 [39] and assumed to be named L2.2.A in a recent review [31]. Our results strengthens the phylogenetical relevance of this sublineage into the global L2 evolutionary history and shows that it was transmitted historically in several Japanese cities. The chronology of emergence of this sublineage relatively to other L2 sublineages was positioned in Shitikov’s L2 diversification scheme. Its position relative to previously described lineages clearly showed that it should be qualified as ancestral according to the current definition of this terminology and diverged from other Beijing sublineages shortly after AAnc1.
Tuberculosis is very ancient in mankind history, however it is still impossible to definitively date its emergence in Asia [10]. Early TB outbreak history in Japan could be related to migrations of people from the 5th century BC to the third century AD [46, 47]. Tuberculosis was known to be present in ancient time in Japan under the name of rōga, that was used in Chinese medicine [82]. A link between this disease and TB, as known in the western medicine, was described in 1857 by a medical doctor, Ōgata Kōan, three years before the opening of Japan [82]. There are many traces in Chinese medical history texts of a disease that can be identified as tuberculosis [83]. The historian Fan Xingzhun gave some clues in his brief study on the history of this disease in China [84]. His analysis of Chinese sources leads him to point out that many terms associated with symptoms suggestive of tuberculosis, appear in ancient sources. The *Classic of Mountains and Seas* (Shanhai Jing, 4th-3rd century BC) describes a remedy to cure the “luo 瘧”, i.e. the scrofulae. However Fan Xingzhun admits that these symptoms may not be specific. He notes that other terms suggestive of the disease appear in other ancient sources such as the *Classic of Poetry* (Shijing, gathering texts spanning from the 11th to the 5th century BC) or in the *Mengzi* (孟子 4th century BC), that Su You 蘇游, a Tang Dynasty (618-907) author, holds as synonyms. Fan Xingzhun stresses that these terms (*zhai 瘴*, *mo 瘢*, *chuanshi 傳尸* (literally: corpse that transmits), *shoubing 瘦病* (literally: thinness disease), *zhuanzhu 轉注*, *fulian 伏練*, *gu zheng 骨蒸* (literally: bones filled with hot steam or hot bones) are often associated with description of states of extreme fatigue and emaciation in early dictionaries and medical books. The “zhai 瘴” ideogram, in particular, that has become one of the most popular to describe, between others, tuberculosis, is defined in the earliest dictionaries as a disease whose main characteristic is weakness, and, later, as a disease that leads the patient “to toss and turn restlessly” until “dying of weakness”. While Fan Xingzhun agrees that all these terms may not necessarily correspond to tuberculosis, the historian nevertheless finds it very likely that the evidence he found in a 14th century book describes a case of pulmonary tuberculosis: “Under the Song (d.960-1127), Shi Dezun, when he was 50 years old got the exhaustion and weight loss disease, he was tossing and turning restlessly until he became very emaciated” [84]. Further, Fan Xingzhun hypothesize (p.97) that tuberculosis prevalence was low in ancient times. However in the *Ancient Book of Tang* (Jiu Tangshu, 舊唐書 941 AD) it is reported that: «under the Wude reign (618-626), in Guanzhong (Shaanxi province), many have the ‘hot bone disease’ 骨蒸», a testimony that argues in favour of a tuberculosis
epidemic. As the term "hot bone disease" (骨蒸) is already mentioned in the medical Canon "The Yellow Emperor's Canon of Medicine" (Huangdi Neijing 黃帝內經, 2BC-2AC), Fan Xingzhun wonders about the possibilities that tuberculosis could have already been epidemic at that time [84].

Going in-depth into the complex historical phylodynamics history of all MTBC lineages was made possible using the TB-Annotator pipeline that analyses more than 50,000 characters including repeated sequences and SNPs [29, 85-89]. With an increasing publicly available number of SRAs, it becomes feasible to disentangle all the threads between ancient and recent historical event that shaped today’s TB pandemia, and to understand its relation to ancient/modern population migrations [10, 90, 91].

Among the set of AAnc5 characteristics we describe a non-synonymous mutation in rpoC. It is well known that rpoC mutations are compensatory mutations mainly found in epidemiologically-successful isolates that also contain specific rpoB mutations [12, 15, 92]. The rpoC mutation could be an adaptative trait that explains the epidemiological success of MDR-TB L2 isolates in outbreaks such as the Central Asian or the B0/W148 outbreak in Russia [6, 93] and the epidemiological success in Georgian prisoners [15]. In the central Asian L2 clade, rpoC mutations may arise within the entire gene and these SNPs are found in epistasia with rpoB mutations [6, 12, 92]. In our study, the rpoC mutation, except for the six most recent last isolates, was found in drug-susceptible isolates; hence it is difficult to consider such a mutation as adaptative or compensatory, but rather as a phylogenetical marker of AAnc5 [15]. Since this SNP was not described in the G3 group from Kobe and Osaka [39], it could be interesting drug-resistance emergence in these two groups of isolates. A recent paper in prisoners in Georgia showed that compensatory mutations and patient incarceration were two independent factors associated with increased transmission, that create a “perfect storm” for MDR-TB transmission [15]. The precise physiological or fitness consequences of this non synonymous rpoC mutation in drug-sensitive isolates was not investigated in this study but it could have also functional consequences [12]. L2 evolution includes the early evolutionary history of AAnc5 and could be linked to yet unknown mutator effects specific of some L2 sublineages [11].

Dating of strains diversification may also give some clues on the social and demographic conditions that fostered past epidemics. Calibration of the molecular clock is
difficult according to samples, time-frames and lineages (between 0.04 to 2.2 SNPs per-genome-per-year) [94]. Dating can be comforted if adequate correlation exists between historical, genomic, epidemiological and demographic facts [91]. In a former similar study, we tested three scenarios to date the MRCA of an L4.2 sublineage occurring in Japan and Turkey, and confronted historical, anthropological, human genetics, paleopathological and genomics results [68]. Here, according to consensus molecular clock of TB on the middle term, the time frame of coalescence of AAnc5 sublineage landmarks could be between 280-310 years ago before present for the closest isolates, and 760-800 years ago for the farthest ones. We based this estimate on L4 and not on L2 mutation rates and hence, are very cautious.

The Tochigi province is famous for its great copper Ashio mine, whose exploitation started beginning of the XVIIth century. One of our estimation of the expansion date of the AAnc 5 is compatible with the opening of the mine in Ashio [95]. Ashio mine could have been one location of AAnc5 expansion and diversification. Looking at the historical spreading dynamics of the AAnc5 group in Japan could be done in the future with the help of japanese investigators by looking at finer gradients of all prevalent L2 ancestral sublineages found in Japan.

Another more recent origin of Japanese AAnc5 could be the importation of ancestral L2 clones by forced workers from Korea or China who were made to work in the mining industry from 1939 till 1945 [82]. Indeed 300,000 Koreans and 38,935 Chinese workers, mostly men were forced to work for Japan during these years [82]. However given the high SNP number accumulated inside AAnc5, this hypothesis seems unlikely. One limitation of this study is that we could not investigate the potential distribution of G3/AAnc5 isolates within all Japan, however our formal proof that G3 and AAnc5 are linked and deeply rooted is a first insight into a complex history. Future study could try to dissecate further, using combination on recent statistics on the total of epidemic cases, in combination to genomic characterization and geographical distribution, the global and the local history of L2 ancestral lineages in Japan [96].

The existence of a relatively high diversity of isolates specific to Tochigi and Kobe and Osaka, reminds of historical transmission restricted to a circumscribed area. Of course, such a picture of an endemic past outbreak is more easily observed in islands settings as was shown recently in New Zealand in Maori people who are hosting a specific “CS” (Colonial S-type) L4.4 Sublineage [91]. Islands are excellent settings to distinguish endemic from imported
species and tuberculosis history is definitively linked to human migration and to local and global demographic history [98]. According to the most recent review on L2 population structure based on more than 5000 genomes, AAnc5 or L2.2.A “is the most basal clade of L2.2.1 and comprises isolates almost entirely from Japan. The deep-branching structure of L2.2.A is suggestive of a previously unrecognized endemic strain” [31]. Since Japan is an island, and a country with a very low current tuberculosis prevalence rate (13 per 100,000 in 2017), it is an excellent setting to identify the historical events linked to past tuberculosis outbreaks [68]. The increasing number of available genomes allows more and more L2 sublineages to be discovered, however their intimate historical and epidemiological relationships remains to be studied in more detail [30, 31].

Conclusions

Thanks to TB-Annotator, a new bioinformatical pipeline that analyzes large amounts of the information contained in SRAs, we mapped an endemic L2 ancestral sublineage from Japan onto the global MTBC phylogeny, and we designated it as asia ancestral 5 (AAnc5), and showed that it was linked to the formerly described G3 group in Kobe and Osaka, and now designated as L2.2.A. This sublineage now appears together with some among the most recent one’s in an unified evolutionary scheme. AAnc5 possesses many specific characters allowing it to be distinct from all other ancestral sublineages described so far in L2. This finding opens new ways of research, to look for the history of L2 in south-east Asia.

Funding

This study did not receive specific funding. The authors declare that they have no conflict of interest.

Data availability statement

All data are publicly available either as SRA Accession numbers on the NCBI or EBI website, or provided herein. The source code of the TB-Annotator pipeline is will be published elsewhere (Senelle et al. in preparation).

Legend to Figures
References

1. WHO. Global Tuberculosis Report 2020. WHO, Geneva, Switzerland, 2020, Oct. 15th. Report No.
2. Devaux I, Kremer K, Heersma H, Van Soolingen D. Clusters of multidrug-resistant Mycobacterium tuberculosis cases, Europe. Emerg Infect Dis. 2009;15(7):1052-60. Epub 2009/07/25. doi: 10.3201/eid1507.080994. PubMed PMID: 19624920; PubMed Central PMCID: PMCPMC2744220.
3. Affolabi D, Faihun F, Sanoussi N, Anyo G, Shamputa IC, Rigouts L, et al. Possible outbreak of streptomycin-resistant Mycobacterium tuberculosis Beijing in Benin. Emerg Infect Dis. 2009;15(7):1123-5. Epub 2009/07/25. doi: 10.3201/eid1507.080697. PubMed PMID: 19624936; PubMed Central PMCID: PMCPMC2744252.
4. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, et al. Evolution and transmission characteristics of Beijing family strains of Mycobacterium tuberculosis in Peru. PLoS One. 2012;7(11):e49651. Epub 2012/11/28. doi: 10.1371/journal.pone.0049651. PubMed PMID: 23185395; PubMed Central PMCID: PMC3939361.
5. Iwamoto T, Grandjean L, Arikawa K, Nakanishi N, Caviedes L, Coronel J, et al. Genetic diversity and transmission characteristics of Beijing family strains of Mycobacterium tuberculosis in Peru. J Clin Microbiol. 2012;50(11):3757-9. Epub 2012/08/31. doi: 10.1128/JCM.02001-12. PubMed PMID: 22933595; PubMed Central PMCID: PMCPMC3486266.
6. Merker M, Barbier M, Cox H, Rasigade JP, Feuerriegel S, Kohl TA, et al. Compensatory evolution drives multidrug-resistant tuberculosis in Central Asia. eLife. 2018;7. Epub 2018/10/31. doi: 10.7554/eLife.38200. PubMed PMID: 30373719; PubMed Central PMCID: PMCPMC6207422.
7. Mokrousov I, Narvskaia O, Vyazovaya A, Otten T, Jiao WW, Gomes LL, et al. Russian "successful" clone B0/W148 of Mycobacterium tuberculosis Beijing genotype: a multiplex PCR assay for rapid detection and global screening. J Clin Microbiol. 2012;50(11):3757-9. Epub 2012/08/31. doi: 10.1128/JCM.02001-12. PubMed PMID: 22933595; PubMed Central PMCID: PMCPMC3486266.
8. Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, et al. Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(26):8136-41. doi: 10.1073/pnas.1424063112. PubMed PMID: 26080405; PubMed Central PMCID: PMCPMC4491734.
9. Liu Q, Luo T, Dong X, Sun G, Liu Z, Gan M, et al. Genetic features of Mycobacterium tuberculosis modern Beijing sublineage. Emerg Microbes Infect. 2016;5:e14. Epub 2016/02/26. doi: 10.1038/emi.2016.14. PubMed PMID: 26905026; PubMed Central PMCID: PMCPMC4777927.
10. Liu Q, Ma A, Wei L, Pang Y, Wu B, Luo T, et al. China's tuberculosis epidemic stems from historical expansion of four strains of Mycobacterium tuberculosis. Nat Ecol Evol. 2018;2(12):1982-92. Epub 2018/11/07. doi: 10.1038/s41559-018-0680-6. PubMed PMID: 30397300; PubMed Central PMCID: PMCPMC6295914.
11. Rad ME, Bifani P, Martin C, Kremer K, Samper S, Rauzier J, et al. Mutations in putative mutator genes of Mycobacterium tuberculosis strains of the W-Beijing family. Emerg Infect Dis. 2003;9(7):838-45.
12. Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, et al. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. Nature genetics. 2011;44(1):106-10. Epub 2011/12/20. doi: 10.1038/ng.1038. PubMed PMID: 22179134; PubMed Central PMCID: PMC3246538.

13. Klote BJ, Kacimi S, Costa-Conceicao E, Gomes HM, Barcellos RB, Panaiotov S, et al. Genomic characterization of MDR/XDR-TB in Kazakhstan by a combination of high-throughput methods predominantly shows the ongoing transmission of L2/Beijing 94-32 central Asian/Russian clusters. BMC Infect Dis. 2019;19(1):553. doi: 10.1186/s12879-019-4201-2. PubMed PMID: 31234780.

14. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nature genetics. 2013;45(7):784-90. Epub 2013/06/12. doi: 10.1038/ng.2656. PubMed PMID: 23749189; PubMed Central PMCID: PMC3776616.

15. Gygli SM, Loiseau C, Jugheli L, Adamia N, Trauner A, Reinhard M, et al. Prisons as ecological drivers of fitness-compensated multidrug-resistant Mycobacterium tuberculosis. Nat Med. 2021. Epub 2021/05/26. doi: 10.1038/s41591-021-01358-x. PubMed PMID: 34031604.

16. Drozin M, Johnson A, Johnson AM. Multidrug resistant tuberculosis in prisons located in former Soviet countries: A systematic review. PLoS One. 2017;12(3):e0174373. Epub 2017/03/24. doi: 10.1371/journal.pone.0174373. PubMed PMID: 28334036; PubMed Central PMCID: PMC5363920.

17. Vidya C, Prabhheesh K. Implications of COVID-19 Pandemic on the Global Trade Networks. EMERGING MARKETS FINANCE AND TRADE. 2020;2020(56):2408-21.

18. Nicita A, Razo C. China: The rise of a trade titan: UNCTAD; 2021 [cited 2021 May 31st, 2021]. Available from: https://unctad.org/news/china-rise-trade-titan.

19. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains. Trends Microbiol. 2002;10(1):45-52. PubMed PMID: 11755085.

20. Mokrousov I, Ly HM, Otten T, Lan NN, Vyshnevskyi B, Hoffner S, et al. Origin and primary dispersal of the Mycobacterium tuberculosis Beijing genotype: clues from human phylogeography. Genome Res. 2005;15(10):1357-64. Epub 2005/09/20. doi: 10.1101/gr.3840605. PubMed PMID: 16169923; PubMed Central PMCID: PMC1240077.

21. Tsolaki AG, Gagneux S, Pym AS, Goguet de la Salmoniere YO, Kreiswirth BN, Van Sooijingen D, et al. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis. J Clin Microbiol. 2005;43(7):3185-91. Epub 2005/07/08. doi: 10.1128/JCM.43.7.3185-3191.2005. PubMed PMID: 16000433; PubMed Central PMCID: PMC1169157.

22. Iwamoto T, Fujiyama R, Yoshida S, Wada T, Shirai C, Kawakami Y. Population structure dynamics of Mycobacterium tuberculosis Beijing strains during past decades in Japan. J Clin Microbiol. 2009;47(10):3340-3. PubMed PMID: 19710282.

23. Mestre O, Luo T, Dos Vultos T, Kremer K, Murray A, Namouchi A, et al. Phylogeny of Mycobacterium tuberculosis Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS One. 2011;6(1):e16020. PubMed PMID: 21283803.

24. Wan K, Liu J, Hauck Y, Zhang Y, Liu J, Zhao X, et al. Investigation on Mycobacterium tuberculosis diversity in China and the origin of the Beijing clade. PLoS One.
2011;6(12):e29190. Epub 2012/01/06. doi: 10.1371/journal.pone.0029190. PubMed PMID: 22220207; PubMed Central PMCID: PMC3248407.

25. Allix-Béguec C, Wahl C, Hanekom M, Nikolayevskyy V, Drobniewski F, Maeda S, et al. Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit-variable-number tandem-repeat loci for subtyping of Mycobacterium tuberculosis Beijing isolates. J Clin Microbiol. 2014;52(1):164-72. Epub 2013/11/01. doi: 10.1128/JCM.02519-13. PubMed PMID: 24172154; PubMed Central PMCID: PMCPMC3911419.

26. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nature genetics. 2015;47(3):242-9. doi: 10.1038/ng.3195. PubMed PMID: 24172154; PubMed Central PMCID: PMCPMC3911419.

27. Shitikov E, Guliaev A, Bespyatykh J, Malakhova M, Kolchenko S, Smirnov G, et al. The role of IS6110 in micro- and macroevolution of Mycobacterium tuberculosis lineage 2. Molecular phylogenetics and evolution. 2019;139:106559. Epub 2019/07/16. doi: 10.1016/j.ympev.2019.106559. PubMed PMID: 31302224.

28. Shitikov E, Kolchenko S, Mokrousov I, Bespyatykh J, Ischenko D, Illina E, et al. Evolutionary pathway analysis and unified classification of East Asian lineage of Mycobacterium tuberculosis. Scientific reports. 2017;7(1):9227. doi: 10.1038/s41598-017-10018-5. PubMed PMID: 28835627; PubMed Central PMCID: PMCPMC5569047.

29. Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, et al. Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. Genome medicine. 2020;12(1):114. Epub 2020/12/16. doi: 10.1186/s13073-020-00817-3. PubMed PMID: 33317631; PubMed Central PMCID: PMCPMC7734807.

30. Freschi L, Vargas R, Jr., Husain A, Kamal SMM, Skrahina A, Tahseen S, et al. Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. Nature communications. 2021;12(1):6099. Epub 2021/10/22. doi: 10.1038/s41467-021-26248-1. PubMed PMID: 34671035; PubMed Central PMCID: PMCPMC8528816.

31. Thawornwattana Y, Mahasirimongkol S, Yanai H, Myat Win Maung H, Cui Z, Chongsuvivatwong V, et al. Revised nomenclature and SNP barcode for Mycobacterium tuberculosis lineage 2. Microbial Genomics. 2021. doi: DOI 10.1099/mgen.0.000697.

32. Srilohasin P, Prammananan T, Paksri K, Phelan JE, Suriyaphol P, Kamolwat P, et al. Genomic evidence supporting the clonal expansion of extensively drug-resistant tuberculosis bacteria belonging to a rare proto-Beijing genotype. Emerg Microbes Infect. 2020;9(1):2632-41. Epub 2020/11/19. doi: 10.1080/22221751.2020.1852891. PubMed PMID: 33205698; PubMed Central PMCID: PMCPMC7738298.

33. Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nature communications. 2014;5:4812. doi: 10.1038/ncomms5812. PubMed PMID: 25176035; PubMed Central PMCID: PMCPMC4166679.

34. Yin QQ, Liu HC, Jiao WW, Li QJ, Han R, Tian JL, et al. Evolutionary History and Ongoing Transmission of Phylogenetic Sublineages of Mycobacterium tuberculosis Beijing Genotype in China. Scientific reports. 2016;6:34353. Epub 2016/09/30. doi: 10.1038/srep34353. PubMed PMID: 27681182; PubMed Central PMCID: PMCPmc5041183.

35. Ajawatanawong P, Yanai H, Smittipat N, Disratthakit A, Yamada N, Miyahara R, et al. A novel Ancestral Beijing sublineage of Mycobacterium tuberculosis suggests the transition site to Modern Beijing sublineages. Scientific reports. 2019;9(1):13718. Epub 2019/09/25. doi: 10.1038/s41598-019-50078-3. PubMed PMID: 31548561; PubMed Central PMCID: PMCPMC6757101.

36. Karmakar M, Trauer JM, Ascher DB, Denholm JT. Hyper transmission of Beijing lineage Mycobacterium tuberculosis: Systematic review and meta-analysis. The Journal of...
37. Casali N, Nikolayevskyy V, Balabanova Y, Ignatyeva O, Kontsevaya I, Harris SR, et al. Microevolution of extensively drug-resistant tuberculosis in Russia. Genome Res. 2012;22(4):735-45. Epub 2012/02/02. doi: 10.1101/gr.128678.111. PubMed PMID: 2294518; PubMed Central PMCID: PMC3317155.

38. Kang HY, Wada T, Iwamoto T, Maeda S, Murase Y, Kato S, et al. Phylogeographical particularity of the Mycobacterium tuberculosis Beijing family in South Korea based on international comparison with surrounding countries. J Med Microbiol. 2010;59(Pt 10):1191-7. Epub 2010/06/26. doi: 10.1099/jmm.0.022103-0. PubMed PMID: 20576748.

39. Wada T, Iwamoto T, Hase A, Maeda S. Scanning of genetic diversity of evolutionarily sequential Mycobacterium tuberculosis Beijing family strains based on genome wide analysis. Infect Genet Evol. 2012;12(7):1392-6. Epub 2012/05/15. doi: 10.1016/j.meegid.2012.04.029. PubMed PMID: 22580240.

40. Park YK, Kang H, Yoo H, Lee SH, Roh H, Kim JJ, et al. Whole-Genome Sequence of Mycobacterium tuberculosis Korean Strain KIT87190. Genome announcements. 2014;2(5). Epub 2014/11/02. doi: 10.1128/genomeA.01103-14. PubMed PMID: 25359915; PubMed Central PMCID: PMC4214991.

41. Wada T, Iwamoto T, Tamaru A, Seto J, Ahiko T, Yamamoto K, et al. Clonality and micro-diversity of a nationwide spreading genotype of Mycobacterium tuberculosis in Japan. PLoS One. 2015;10(3):e0118495. Epub 2015/03/04. doi: 10.1371/journal.pone.0118495. PubMed PMID: 25734518; PubMed Central PMCID: PMCPMC4348518.

42. Mizukoshi F, Miyoshi-Akiyama T, Iwai H, Suzuki T, Kiritani R, Kirikae T, et al. Genetic diversity of Mycobacterium tuberculosis isolates from Tochigi prefecture, a local region of Japan. BMC Infect Dis. 2017;17(1):365. Epub 2017/05/27. doi: 10.1186/s12879-017-2457-y. PubMed PMID: 28545488; PubMed Central PMCID: PMCPMC5445273.

43. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(14):4871-6.

44. Mokrousov I. Genetic geography of Mycobacterium tuberculosis Beijing genotype: a multifaceted mirror of human history? Infect Genet Evol. 2008;8(6):777-85. doi: 10.1016/j.meegid.2008.07.003. PubMed PMID: 18691674.

45. Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmacari SK, Calacal GC, et al. Mapping human genetic diversity in Asia. Science. 2009;326(5959):1541-5. Epub 2009/12/17. doi: 10.1126/science.1177074. PubMed PMID: 2007900.

46. Suzuki T, Fujita H, Choi JG. Brief communication: new evidence of tuberculosis from prehistoric Korea-Population movement and early evidence of tuberculosis in far East Asia. American journal of physical anthropology. 2008;136(3):357-60. doi: 10.1002/ajpa.20811. PubMed PMID: 18322918.

47. Suzuki T, Inoue T. Earliest Evidence of Spinal Tuberculosis from the Anoelithic Yayoi Period in Japan. International Journal of Osteoarchaeology. 2007;17:392-492.

48. Wada T, Iwamoto T. Allelic diversity of variable number of tandem repeats provides phylogenetic clues regarding the Mycobacterium tuberculosis Beijing family. Infect Genet Evol. 2009;9(5):921-6. Epub 2009/06/23. doi: 10.1016/j.meegid.2009.06.012. PubMed PMID: 19540936.

49. Yokoyama E, Hachisu Y, Hashimoto R, Kishida K. Concordance of variable-number tandem repeat (VNTR) and large sequence polymorphism (LSP) analyses of Mycobacterium tuberculosis strains. Infect Genet Evol. 2010;10(7):913-8. doi: 10.1016/j.meegid.2010.05.013. PubMed PMID: 20561911.
50. Yokoyama E, Hachisu Y, Hashimoto R, Kishida K. Population genetic analysis of Mycobacterium tuberculosis Beijing subgroup strains. Infect Genet Evol. 2012;12(4):630-6. doi: 10.1016/j.meegid.2011.10.016. PubMed PMID: 22041504.
51. Nakanishi N, Wada T, Arikawa K, Millet J, Rastogi N, Iwamoto T. Evolutionary robust SNPs reveal the misclassification of Mycobacterium tuberculosis Beijing family strains into sublineages. Infect Genet Evol. 2013;16C:174-7. Epub 2013/02/27. doi: 10.1016/j.meegid.2013.02.007. PubMed PMID: 23438651.
52. Yokoyama E, Hachisu Y, Iwamoto T, Nakanishi N, Arikawa K, Wada T, et al. Comparative analysis of Mycobacterium tuberculosis Beijing strains isolated in three remote areas of Japan. Infect Genet Evol. 2015;34:444-9. Epub 2015/06/23. doi: 10.1016/j.meegid.2015.06.019. PubMed PMID: 26096775.
53. Iwai H, Kato-Miyazawa M, Kirikae T, Miyoshi-Akiyama T. CASTB (the comprehensive analysis server for the Mycobacterium tuberculosis complex): A publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. Tuberculosis (Edinb). 2015;95(6):843-4. doi: 10.1016/j.tube.2015.09.002. PubMed PMID: 26542225.
54. Guyeux C, Sola C, Nous C, Refregier G. CRISPRbuilder-TB: "CRISPR-builder for tuberculosis". Exhaustive reconstruction of the CRISPR locus in mycobacterium tuberculosis complex using SRA. PLoS computational biology. 2021;17(3):e1008500. Epub 2021/03/06. doi: 10.1371/journal.pcbi.1008500. PubMed PMID: 33667225; PubMed Central PMCID: PMCPMC7968741.
55. 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. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(18):9869-74. PubMed PMID: 9275218.
56. Malm S, Linguissi LS, Tekwu EM, Vouyoungui JC, Kohl TA, Beckert P, et al. New Mycobacterium tuberculosis Complex Sublineage; Brazzaville, Congo. Emerg Infect Dis. 2017;23(3):423-9. doi: 10.3201/eid2303.160679. PubMed PMID: 28221129; PubMed Central PMCID: PMCPMC5382753.
57. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbon MH, Bobadilla del Valle M, et al. Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. Journal of bacteriology. 2006;188(2):759-72. PubMed PMID: 16385065.
58. Borrell S, Trauner A, Brites D, Rigouts L, Loiseau C, Coscolla M, et al. Reference set of Mycobacterium tuberculosis clinical strains: A tool for research and product development. PLoS One. 2019;14(3):e0214088. Epub 2019/03/26. doi: 10.1371/journal.pone.0214088. PubMed PMID: 30908506; PubMed Central PMCID: PMCPMC6433267.
59. Faksrri K, Xia E, Tan JH, Teo YY, Ong RT. In silico region of difference (RD) analysis of Mycobacterium tuberculosis complex from sequence reads using RD-Analyzer. BMC genomics. 2016;17(1):847. Epub 2016/11/04. doi: 10.1186/s12864-016-3213-1. PubMed PMID: 27806686; PubMed Central PMCID: PMCPMC5093977.
60. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eigelmeier K, et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(6):3684-9. Epub 2002/03/14. doi: 10.1073/pnas.052548299. PubMed PMID: 11891304; PubMed Central PMCID: PMCPMC122584.
61. Palittapongarnpim P, Ajawatanawong P, Viratynosin W, Smittipat N, Disratthakit A, Mahasirimongkol S, et al. Evidence for Host-Bacterial Co-evolution via Genome Sequence Analysis of 480 Thai Mycobacterium tuberculosis Lineage 1 Isolates. Scientific reports.
62. Coscolla M, Brites D, Menardo F, Loiseau C, Borrell S, Otchere ID, et al. Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history. bioRxiv [Internet]. 2020. Available from: https://doi.org/10.1101/2020.06.10.141788.

63. 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. Nature genetics. 2016;48(12):1535-43. doi: 10.1038/ng.3704. PubMed PMID: 27798628.

64. Homolka S, Projahn M, Feuerriegel S, Ubben T, Diel R, Nubel U, et al. High Resolution Discrimination of Clinical Mycobacterium tuberculosis Complex Strains Based on Single Nucleotide Polymorphisms. PLoS One. 2012;7(7):e39855. Epub 2012/07/07. doi: 10.1371/journal.pone.0039855. PubMed PMID: 22768315; PubMed Central PMCID: PMC3388094.

65. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS One. 2009;4(11):e7815. Epub 2009/11/17. doi: 10.1371/journal.pone.0007815. PubMed PMID: 19915672; PubMed Central PMCID: PMC2772813.

66. 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. Nature genetics. 2013;45(10):1176-82. Epub 2013/09/03. doi: 10.1038/ng.2744. PubMed PMID: 23995134.

67. Mokrousov I, Shitikov E, Skiba Y, Kolchenko S, Chernyaeva E, Vyazovaya A. Emerging peak on the phylogeographic landscape of Mycobacterium tuberculosis in West Asia: Definitely smoke, likely fire. Molecular phylogenetics and evolution. 2017;116:202-12. doi: 10.1016/j.ympev.2017.09.002. PubMed PMID: 28893611.

68. Refregier G, Abadia E, Matsumoto T, Ano H, Takashima T, Tsuyuguchi I, et al. Turkish and Japanese Mycobacterium tuberculosis sublineages share a remote common ancestor. Infect Genet Evol. 2016;45:461-73. doi: 10.1016/j.meegid.2016.10.009. PubMed PMID: 27746295.

69. Tantivitayakul P, Juthayothin T, Ruangchai W, Smittipat N, Disratthakit A, Mahasirimongkol S, et al. Identification and in silico functional prediction of lineage-specific SNPs distributed in DosR-related proteins and resuscitation-promoting factor proteins of Mycobacterium tuberculosis. Heliyon. 2020;6(12):e05744. Epub 2020/12/29. doi: 10.1016/j.heliyon.2020.e05744. PubMed PMID: 33364506; PubMed Central PMCID: PMC7753917.

70. Ates LS, Dippenaar A, Sayes F, Pawlik A, Bouchier C, Ma L, et al. Unexpected Genomic and Phenotypic Diversity of Mycobacterium africanum Lineage 5 Affects Drug Resistance, Protein Secretion, and Immunogenicity. Genome biology and evolution. 2018;10(8):1858-74. Epub 2018/07/17. doi: 10.1093/gbe/evy145. PubMed PMID: 30010947; PubMed Central PMCID: PMC6071665.

71. Lipworth S, Jajou R, de Neeling A, Bradley P, van der Hoek W, Maphalala G, et al. SNP-IT Tool for Identifying Subspecies and Associated Lineages of Mycobacterium tuberculosis Complex. Emerg Infect Dis. 2019;25(3):482-8. Epub 2019/02/23. doi: 10.3201/eid2503.180894. PubMed PMID: 30789126; PubMed Central PMCID: PMC6390766.

72. Kohl TA, Utpatel C, Schleusener V, De Filippo MR, Beckert P, Cirillo DM, et al. MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium
tuberculosis complex isolates. PeerJ. 2018;6:e5895. Epub 2018/11/28. doi: 10.7717/peerj.5895. PubMed PMID: 30479891; PubMed Central PMCID: PMCPMC6238766.

73. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nature genetics. 2018;50(2):307-16. doi: 10.1038/s41588-017-0029-0. PubMed PMID: 29358649.

74. Coker OO, Chaiprasert A, Ngamphiw C, Tongsima S, Regmi SM, Clark TG, et al. Genetic signatures of Mycobacterium tuberculosis Nonthaburi genotype revealed by whole genome analysis of isolates from tuberculous meningitis patients in Thailand. PeerJ. 2016;4:e1905. Epub 2016/04/27. doi: 10.7717/peerj.1905. PubMed PMID: 27114869; PubMed Central PMCID: PMCPMC4841212.

75. Koster K, Largen A, Foster JT, Drees KP, Qian L, Desmond EP, et al. Whole genome SNP analysis suggests unique virulence factor differences of the Beijing and Manila families of Mycobacterium tuberculosis found in Hawaii. PLoS One. 2018;13(7):e0201146. Epub 2018/07/24. doi: 10.1371/journal.pone.0201146. PubMed PMID: 30036392; PubMed Central PMCID: PMCPMC6056056.

76. Couvin D, David A, Zozio T, Rastogi N. Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the Mycobacterium tuberculosis genotyping database. Infect Genet Evol. 2019;72:31-43. Epub 2018/12/30. doi: 10.1016/j.meegid.2018.12.030. PubMed PMID: 30593925.

77. Refrégier G, Sola C, Guyeux C. Unexpected diversity of CRISPR unveils some evolutionary patterns of repeated sequences in Mycobacterium tuberculosis. BMC genomics. 2020;21:841. doi: https://doi.org/10.1186/s12864-020-07178-6.

78. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3. Epub 2014/01/24. doi: 10.1093/bioinformatics/btu033. PubMed PMID: 24451623; PubMed Central PMCID: PMCPMC3998144.

79. Kloepper TH, Huson DH. Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC evolutionary biology. 2008;8:22. Epub 2008/01/26. doi: 10.1186/1471-2148-8-22. PubMed PMID: 18218099; PubMed Central PMCID: PMCPMC2253509.

80. Han SJ, Song T, Cho YJ, Kim JS, Choi SY, Bang HE, et al. Complete genome sequence of Mycobacterium tuberculosis K from a Korean high school outbreak, belonging to the Beijing family. Stand Genomic Sci. 2015;10:78. Epub 2015/10/17. doi: 10.1186/s40793-015-0071-4. PubMed PMID: 26473025; PubMed Central PMCID: PMCPMC4606834.

81. Alix-Beguec C, Harmsen D, Weniger D, Supply P, Niemann S. Evaluation and user-strategy of MIRU-VNTRplus, a multifunctional database for on-line analysis of genotyping data and phylogenetic identification of Mycobacterium tuberculosis complex isolates. J Clin Microbiol. 2008;46(8):2692-9. PubMed PMID: 18550737.

82. Thomann B. L'hygiène nationale, la société civile et la reconnaissance de la silicose comme maladie professionnelle au Japon (1868-1960). Revue d'Histoire Moderne et Contemporaine. 2009;2009(156-1):142-76.

83. Barberis I, Bragazzi NL, Galluzzo L, Martini M. The history of tuberculosis: from the first historical records to the isolation of Koch's bacillus. J Prev Med Hyg. 2017;58(1):E9-E12. Epub 2017/05/19. PubMed PMID: 28515626; PubMed Central PMCID: PMCPMC5432783.

84. Fan X. Zhongguo bingshi xinyi 中國病史新義 (New meaning of disease in Chinese history). Beijing: Zhongyi guji chuban; 1989.

85. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting:
recommendations for a standardized methodology. J Clin Microbiol. 1993;31(2):406-9. PubMed PMID: 8381814.
86. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol. 1997;35(4):907-14. PubMed PMID: 9157152.
87. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. J Clin Microbiol. 2006;44(12):4498-510. Epub 2006/09/29. doi: 10.1128/JCM.01392-06. PubMed PMID: 17005759; PubMed Central PMCID: PMC398431.
88. Freschi L, Vargas R, Hussain A, Kamal SMM, Skrahina A, Tahseen S, et al. Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. bioRxiv. 2020:2020.09.29.293274. doi: 10.1101/2020.09.29.293274.
89. Menardo F, Rutaihwa L, Zwyer M, Comas I, Conceição E, et al. Local adaptation in populations of Mycobacterium tuberculosis endemic to the Indian Ocean Rim [version 2; peer review: 2 approved]. PLoS Pathog. 2021;10(60). doi: 10.12688/f1000research.28318.1.
90. Pepperell CS, Granka JM, Alexander DC, Behr MA, Chui L, Gordon J, et al. Dispersal of Mycobacterium tuberculosis via the Canadian fur trade. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(16):6526-31. Epub 2011/04/06. doi: 10.1073/pnas.1016708108. PubMed PMID: 21464295; PubMed Central PMCID: PMCPMC3080970.
91. Mulholland CV, Shockey AC, HA; Cursons RT, O'Toole RF, Gautam SS, et al. Dispersal of Mycobacterium tuberculosis driven by historical european trade in the South Pacific. Frontiers in microbiology. 2019;doi: 10.3389/fmicb.2019.02778. Epub 04 December 2019.
92. de Vos M, Muller B, Borrell S, Black PA, van Helden PD, Warren RM, et al. Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. Antimicrob Agents Chemother. 2013;57(2):827-32. Epub 2012/12/05. doi: 10.1128/AAC.01541-12. PubMed PMID: 23208709; PubMed Central PMCID: PMC3553702.
93. Li QJ, Jiao WW, Yin QQ, Xu F, Li JQ, Sun L, et al. Compensatory Mutations of Rifampin Resistance Are Associated with Transmission of Multidrug-Resistant Mycobacterium tuberculosis Beijing Genotype Strains in China. Antimicrob Agents Chemother. 2016;60(5):2807-12. Epub 2016/02/24. doi: 10.1128/AAC.02358-15. PubMed PMID: 26902762; PubMed Central PMCID: PMC4862491.
94. Menardo F, Duchene S, Brites D, Gagneux S. The molecular clock of Mycobacterium tuberculosis. PLoS pathogens. 2019;15(9):e1008067. Epub 2019/09/13. doi: 10.1371/journal.ppat.1008067. PubMed PMID: 31513651.
95. Wikipedia. Ashio Copper mine https://en.wikipedia.org/wiki/Ashio_Copper_Mine: Wikipedia, 2021 [cited 2021 May 31st 2021]. Available from: https://en.wikipedia.org/wiki/Ashio_Copper_Mine.
96. Ota M, Hoshino Y, Hirao S. Analysis of 605 tuberculosis outbreaks in Japan, 1993–2015: time, place and transmission site. Epidemiology and infection. 2021;149:e85. Epub 2021/03/22. doi: 10.1017/S0950268821000625.
97. Sola C, Ferdinand S, Sechi LA, Zanetti S, Martial D, Mammina C, et al. Mycobacterium tuberculosis molecular evolution in Western Mediterranean Islands of Sicily and Sardinia. Infect Genet Evol. 2005;5(2):145-56.