MICROBIAL GENOMICS

RESEARCH ARTICLE

Thawornwattana *et al.*, *Microbial Genomics* 2021;7:000697

DOI 10.1099/mgen.0.000697

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Revised nomenclature and SNP barcode for *Mycobacterium tuberculosis* lineage 2

Yuttapong Thawornwattana[1,2], Surakameth Mahasirimongkol[3], Hideki Yanai[4], Htet Myat Win Maung[5,6], Zhezhe Cui[6,7], Virasakdi Chongsuvivatwong[6] and Prasit Palittapongarnpim[1,8,*]

Abstract

Mycobacterium tuberculosis (Mtb) lineage 2 (L2) strains are present globally, contributing to a widespread tuberculosis (TB) burden, particularly in Asia where both prevalence of TB and numbers of drug resistant TB are highest. The increasing availability of whole-genome sequencing (WGS) data worldwide provides an opportunity to improve our understanding of the global genetic diversity of Mtb L2 and its association with the disease epidemiology and pathogenesis. However, existing L2 sublineage classification schemes leave >20% of the Modern Beijing isolates unclassified. Here, we present a revised SNP-based classification scheme of L2 in a genomic framework based on phylogenetic analysis of >4000 L2 isolates from 34 countries in Asia, Eastern Europe, Oceania and Africa. Our scheme consists of over 30 genotypes, many of which have not been described before. In particular, we propose six main genotypes of Modern Beijing strains, denoted L2.2.M1–L2.2.M6. We also provide SNP markers for genotyping L2 strains from WGS data. This fine-scale genotyping scheme, which can classify >98% of the studied isolates, serves as a basis for more effective monitoring and reporting of transmission and outbreaks, as well as improving genotype-phenotype associations such as disease severity and drug resistance. This article contains data hosted by Microreact.

DATA SUMMARY

(1) Raw *M. tuberculosis* sequence data are available in the European Nucleotide Archive (ENA) database, and can be accessed with accession numbers provided in Table S2 (available in the online version of this article). A list of sources of isolates from previous studies and references is provided in Table S1.

(2) *M. tuberculosis* strain H37Rv is available from GenBank with accession number NC_000962.3.

(3) Script used for variant calling is available at https://github.com/CENMIG/snpplet.

(4) A list of genotype-specific SNPs and a list of barcoding SNPs are available at Figshare: https://doi.org/10.6084/m9.figshare.14709513.v1 [1]. The full list with additional information is in Table S7 and the barcoding SNPs are summarized in Table S8.

(5) Vcf and multiple sequence alignment files are available at Figshare.

(6) Interactive phylogenetic trees are available online at Microreact for the discovery set (https://microreact.org/project/4P2iPeBx1Y66TyJfojNM1o) and the discovery +test set (https://microreact.org/project/4kJjFVPWe3TbfwFtqY3ED8)

INTRODUCTION

Mycobacterium tuberculosis (Mtb) lineage 2 (L2) is a one of the major global strains, with high prevalence in Asia [2].

Received 05 March 2021; Accepted 24 September 2021; Published 17 November 2021

Author affiliations: [1]Pornchai Matangkasombut Center for Microbial Genomics, Department of Microbiology, Faculty of Science, Mahidol University, Bangkok 10400, Thailand; [2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA; [3]Department of Medical Sciences, Ministry of Public Health, Nonthaburi 11000, Thailand; [4]Fukujuji Hospital and Research Institute of Tuberculosis, Japan Anti-Tuberculosis Association, Kiyose 204-8533, Japan; [5]National TB Control Programme, Department of Public Health, Ministry of Health and Sports, Naypyitaw 15011, Myanmar; [6]Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Had Yai 90110, Thailand; [7]Department of Tuberculosis Control, Guangxi Zhuang Autonomous Region Center for Disease Control and Prevention, Nanning, Guangxi, 530028, PR China; [8]National Science and Technology Development Agency, Pathumthani 12120, Thailand.

***Correspondence:** Prasit Palittapongarnpim, Prasit.pal@mahidol.ac.th

Keywords: Beijing strain; genotyping; Lineage 2; *Mycobacterium tuberculosis*; phylogeny; whole genome sequencing.

Abbreviations: AA, Asia Ancestral; CAO, Central Asia Outbreak; L2, lineage 2; LSP, large sequence polymorphism; MDR, multidrug resistant; Mtb, Mycobacterium tuberculosis; RD, region of difference; SIT, spoligotype international type; SNP, single nucleotide polymorphism; VNTR, variable number tandem repeat; WGS, whole-genome sequencing.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Nine supplementary tables, supplemetary text and nine supplementary figures are available with the online version of this article.

It belongs to a species complex that comprises eight main human-adapted Mtb lineages. Mtb L2 likely originated in a region around southern East Asia or Southeast Asia [3–5], expanded in China and spread to the rest of the world, particularly in Central Asia, Eastern Europe and East Africa [2, 3, 6]. L2 strains have been associated with higher virulence [7, 8], increased transmissibility [9–11] and high prevalence of multidrug resistance [12, 13]. However, global genomic diversity and population structure of Mtb L2 remain poorly characterized. This hampers reporting and monitoring of local and global transmission and outbreaks, and meaningful associations of sublineages with epidemiologically or clinically important phenotypes such as drug resistance, pathogenicity, virulence, disease manifestations and clinical outcomes.

Several genotyping methods and classification schemes exist for Mtb lineages, including large sequence polymorphism (LSP) typing [14], VNTR typing [15], spoligotyping [16] and single nucleotide polymorphism (SNP) typing based on a limited set of genes [17–19]. However, these schemes tend not to have sufficient resolution to differentiate epidemiologically important or emerging L2 strains at a sublineage level. This is partly due to relatively limited genetic diversity of L2 strains compared with other Mtb strains [9, 20]. For instance, the LSP scheme for L2 is based on the presence/absence of five large genomic deletions (RD105, RD207, RD181, RD150 and RD142) [14]. This deletion pattern can be used to differentiate three major sublineages of L2, namely L2.1 (only RD105 deleted), L2.2.2 (RD105 and RD207 deleted) and L2.2.1 (RD105, RD207 and RD181 deleted). RD150 and RD142 deletions are unique to two small subgroups of L2.2.1. L2.1 and L2.2 are also known as proto-Beijing and Beijing strains, respectively [21, 22]. Spoligotyping, a widely-used cost-effective genotyping method for Mtb isolates, has a low discriminatory power for L2 strains: spoligotypes of L2 can only distinguish primarily between L2.1 (mainly SIT523 (Manu Ancestor): 777777777777771) and L2.2 (mainly SIT1 (Beijing): 000000000003771). SNP-based genotyping schemes are based on the presence of several selected SNPs. The most widely-used SNPs for L2 genotyping are probably the ones used to differentiate between Ancestral and Modern Beijing strains [17]. The Modern Beijing strains were first described based on insertion profiles of a transposable element called IS6110 [23]. This group can be characterised by an IS6110 insertion pattern in the NTF region [24, 25] as well as allelic variants of genes coding for DNA repair enzymes called *mutT2*, *ogt* and *mutT4* [17]. More comprehensive differentiation schemes were also proposed, which classified Mtb L2 into several ST (sequence type) [18] or Bmyc [19]. Although occasionally useful, the limited numbers of informative SNPs used by those schemes resulted in limited discrimination powers and genotypes, which do not always correspond to valid phylogenetic clades.

Limitations of traditional genotyping methods for L2 strains motivate higher resolution typing methods based on whole-genome sequencing (WGS) data. For instance, genome-wide SNP-based genotyping relies on SNPs shared exclusively within each genotype. These genotype-specific SNPs can be identified from genome-wide SNPs extracted from WGS data in a phylogenetic framework. Several SNP-based genotyping schemes for L2 strains have been proposed [5, 22, 26–28]. For instance, Merker *et al.* [5] proposed a classification scheme that splits L2.2 (Beijing) into eight major genotypes based on a global collection of ~100 isolates, namely Asia Ancestral 1–3, Asian African 1–2, Pacific RD150, Europe-Russia B0/W148 and Central Asian. Liu *et al.* [27] proposed a set of 44 SNPs for distinguishing the Modern Beijing strains within L2.2.1. Some authors refer to the Modern Beijing strains as L2.3 [10, 21] but this terminology is avoided here as it renders L2.2 paraphyletic. Two additional genotypes have been recently proposed, Asian African 3 [26] and Asia Ancestral 4 [29], bringing the total number of major L2 sublineages to twelve. Rutaihwa *et al.* [30] proposed an alternative nomenclature system whereby L2.2 comprises ten genotypes L2.2.1–L2.2.10 for Asia Ancestral 1–3, Asian African 1, Asian African 3, Pacific RD150, Asian African 2, RD142 (nested within Asian African 2), Europe-Russia B0/W148 and Central Asian, respectively. Note that L2.2.1 and L2.2.2 in this scheme are different from the previous definition of these two genotypes. However, existing SNP-based genotyping schemes are either based on a small collection of L2 isolates or based on isolates from limited geographical regions and are not representative of the global genetic diversity of L2. In particular, isolates from endemic regions such as East Asia and Southeast Asia tend to be under-represented. Thus, they may not adequately capture the diversity of L2 strains in endemic areas for use in epidemiological applications such as local transmission monitoring. In fact, a large proportion of L2 strains remain unclassified at a sublineage level beyond L2.2.1 by any existing genotyping scheme, particularly Modern Beijing strains [29].

Here, we analysed a collection of over 4,000 Mtb L2 genomes from Asia, Eastern Europe, Oceania and East Africa to gain insights into the global genetic diversity and population

Impact Statement

Mycobacterium tuberculosis lineage 2 (L2) is widespread globally and may be expanding in several countries. Genotypic classification is crucial for better understanding of the relationships between the genetic and phenotypic characteristics of the bacteria. Several genotypic classification schemes have been proposed. Unfortunately, many L2 isolates were still left unclassified beyond the level of Ancestral or Modern Beijing strains. The classification scheme proposed here can classify 98% of the L2 isolates and is kept as backward compatible as possible. The scheme is useful for identification of phenotypic characteristics that could be associated with genotypes, recognition of genetic relationships between isolates in different countries. The scheme is also a basis for further studies in pathogenesis such as identification of interesting homoplastic SNPs that are resulted from convergent evolution.

structure of L2 strains. The selected areas cover likely places of origin of Mtb L2 as well as its major sublineages and the Modern Beijing strains. Using a phylogenetic framework, we propose a hierarchical classification and nomenclature system that captures most of the diversity of L2 strains sampled to date, keeping backward-compatibility with existing major genotyping schemes as much as possible. We also provide SNP markers for genotyping existing and new sublineages of L2 strains. Our proposed scheme is expected to greatly facilitate identification and reporting of sublineage-level L2 strains, e.g. in monitoring transmission and outbreak investigations, particularly in endemic regions. It also serves as a foundation for studying evolution, epidemiology and pathogenesis of Mtb L2.

METHODS

Whole-genome sequence data and variant calling

We compiled a collection of whole-genome sequencing data of 4,425 Mtb L2 isolates from countries in Asia, Eastern Europe, Oceania and Africa with a particular emphasis on isolates from East Asia, Southeast Asia and South Asia where L2 is endemic (Table S2). All sequencing data were from the Illumina platform and from over 50 studies. Studies focusing on large outbreaks or highly clonal strains were not included.

Raw sequence data in the fastq format were downloaded from the European Nucleotide Archive (ENA) for each study (https://www.ebi.ac.uk/ena). The short reads were trimmed to remove adapter sequences and low quality read positions using trimmomatic v0.39 (sliding-window trimming with window size of 4 and read quality threshold of 30) [31]. Depending on the quality of the raw reads which differ among studies, variation of the trimming procedure or parameter values may be used. The trimmed reads were then mapped to the H37Rv reference genome (NC_000962.3) using bwa mem [32]. Picard's MarkDuplicates was used to identify duplicate reads before per-sample variant calling using GATK HaplotypeCaller in a haploid model [33], excluding bases with a quality score below 20.

We performed the following four sample quality checks to exclude isolates with poor quality sequence/genotype data, or were redundant. **(1) Read-mapping coverage:** A sample was excluded if it had a median depth of mapped reads below 10 or median breadth of coverage below 10%, where at least 20 reads were required for a position to be counted. We also performed parallel analyses using more stringent criteria of median cutoff of 20 and median breadth (at depth 20) of 50%. **(2) Identity:** For multiple isolates with identical BioSample accession number, only the sample with the highest median depth of mapped reads was included. For studies where multiple clinical samples were taken from a single patient at different times, only one sample per patient was retained (the earliest sample collected was used if the information was available). **(3) Contamination:** Samples with the mean absolute difference between the observed and expected per-read GC content distributions greater than 30% (based on fastqc) were excluded. **(4) Mixed strains:** Samples that were not L2 or had SNPs indicating the presence of multiple non-nested genotypes were excluded. We identified strains using SNP markers from several schemes (see next section). A sample was identified as mixed strains if it had more than one genotype-specific SNP for at least two different genotypes, or had genotype-specific SNPs for more than two non-nested genotypes. Note that SNPs from Coll *et al.* [22] appeared to contain a small proportion of SNPs not specific to a genotype for most genotypes in our data, so this scheme by itself was not useful for identifying mixed strains. The sample quality control step resulted in 4,425 L2 isolates.

To compare SNPs across samples, we performed joint genotyping of all samples using GATK GenotypeGVCFs [33], using per-sample variant calls as inputs. We excluded indels and low-quality SNPs (filter: QD <2 or MQ <40). To minimize false positive variant calls, SNPs located within repeat regions, regions annotated as mobile element, phage, IS6110 or 13E12, PE/PPE genes and known drug resistant genes were excluded as previously described [29]. This filtering resulted in 140,049 high-quality SNPs, which were annotated with SnpEff [34] using annotation from the H37Rv reference genome. This SNP set was then converted into a multiple sequence alignment using a custom shell script. Both point deletions and uncalled variants were converted into gaps in the alignment.

Strain typing schemes

We implemented the following six SNP-typing schemes for identifying strains in our collection and for comparing with our proposed scheme. (1) Ajawatanawong *et al.* [29] scheme derived from >1,000 isolates from Chiang Rai, Thailand, representing local diversity of lineages 1–4 in the region. (2) Coll *et al.* [22] scheme derived from a global collection but had limited resolution for L2 strains. (3) Shitikov *et al.* [26] provided one of the most comprehensive classification schemes for L2 strains, compiled from several existing schemes, and can identify most of the known L2 genotypes including eight major sublineages of L2.2.1 (Asia Ancestral 2–3, Asian African 1–3, Pacific RD150, Europe/Russia B0/W187 and Central Asian) as well as two outbreak strains of the Central Asian clade: Clade A and Central Asia Outbreak (CAO). (4) Liu *et al.* [27] provided 44 SNPs specific to Modern Beijing strains. (5) South African strain of Asia Ancestral 1 (AA1SA) [35]. (6) Mestre *et al.* [19] Bmyc scheme is based on the presence/absence pattern of SNPs in genes involved in replication, recombination and repair. Bmyc genotypes are not necessarily monophyletic. We also performed *in silico* spoligotyping using two methods: SpoTyping v2.1 [36] and Galru v1.0.0 [37].

Phylogenetic analysis

A maximum-likelihood (ML) phylogenetic tree of 4,425 isolates was inferred using IQ-TREE v2 [38] with ultrafast bootstrap supports from 1,000 replications. The best-fit nucleotide substitution model was GTR+G4 as determined by ModelFinder [39]. The lineage 4 H37Rv reference strain (GenBank accession number NC_000962.3) was used as an outgroup for rooting the tree.

We also identified genomic clusters as clades in the phylogeny containing isolates that can be linked via pairwise SNP distances of at most 12 [40, 41]. A cluster may contain pairs of isolates that differed by more than 12 SNPs.

Identification of clades and genotype assignment

We aimed to assign genotypes to previously unclassified ancestral and modern Beijing strains while keeping backward compatibility with existing genotyping schemes as much as possible. We required that a genotype must satisfy the following four criteria. (a) Isolates that have the same genotype must form a monophyletic clade in the phylogenetic tree. (b) The clade must have a bootstrap support at least 90% at the branch leading to the root node of the clade. (c) All isolates within the clade must share at least one common SNP that is different from the variant of the outgroup. (d) A genotype must be represented by at least 10 isolates for Ancestral Beijing genotypes, and at least 20 isolates for Modern Beijing genotypes. This difference in the sample size cutoff reflected the fact that Modern Beijing strains are much more genetically similar to each other (~200 SNP differences between genotypes on average) compared with Ancestral Beijing strains (~300–500 SNP differences between genotypes on average).

Population diversity and divergence

Once the genotypes have been defined, we calculated three summary statistics of genetic diversity within and between genotypes. First, a genome-wide average of nucleotide diversity was calculated as the proportion of pairwise differences in allelic types among all pairs of isolates within each genotype. Second, the mean pairwise SNP distance among isolates from the same genotype compared with the distance where one isolate is from other genotype. Third, the average F_{ST} for each pair of genotypes based on the Hudson estimator [42]. We also performed principal component analysis (PCA) to illustrate clustering of genotypes. Allele frequency data were centred and scaled to unit variance. All computations were performed using scikit-allele v1.3.1 [43].

Identification of genotype-specific SNPs and genotyping SNPs

For each genotype, we identified SNPs that were shared only by samples within the genotype using a custom python script. We then selected a few genotype-specific SNPs as stable markers for genotyping. We preferred synonymous SNPs in T cell epitopes, in essential genes or in the third codon position (in order of preference) as well as those used by previous barcoding schemes [22, 26]. For genotypes with no more than two clade-specific SNPs, all SNPs were used. If multiple SNPs satisfied all those properties, we retained a few. For genotypes with no SNPs satisfying those preferable properties, we relaxed the criteria and considered other types of variants and states of essentiality [44].

The T cell epitopes in Mtb have been shown to be hyperconserved in Mtb [45]. A list of T cell epitopes was obtained from the Immune Epitope Database (IEDB) (https://www.iedb.org/) on 28 August 2020 [46]. We initially retrieved 2116 T cell epitopes using the following selection criteria: linear epitope, organism: *Mycobacterium tuberculosis* (ID: 1773), positive assays only, T cell assays, any MHC restriction, human host, any disease and any reference type. After excluding epitopes that cannot be assigned to a unique genomic location of the H37Rv reference strain, we obtained a final list of 1720 epitopes from 312 annotated ORFs. However, only three SNPs fall within a T-cell epitope, this property was less useful for selecting barcoding SNPs.

We also annotated variants within an ORF using four states of essentiality: essential (ES), growth defect (GD), nonessential (NE) and growth advantage (GA) [44].

Validation of the genotyping scheme

We validated the proposed genotyping scheme on an independent dataset of 1,207 publicly available isolates of Mtb L2 downloaded from the NCBI's Sequence Read Archive (SRA) database (Table S2). The raw sequence data were processed in the same way as for the main (discovery) dataset (n=4,425). Joint genotyping of all samples of all 5,632 was performed using GATK GenotypeGVCFs [33] to produce a multi-sample vcf file which was then converted to multiple sequence alignment. An ML phylogenetic tree was inferred under the GTR+G4 model using IQ-TREE v2 [38]. We assigned a genotype to the isolates using genotype-specific SNPs from Table S7. The ML tree was used to confirm the accuracy of the assigned genotypes: a genotype assignment of an isolate was correct if the isolate fell within a well-supported (bootstrap support >90%) monophyletic clade of isolates with the same genotype.

RESULTS

Strain diversity and the phylogeny of Mtb L2

To characterize the genetic diversity and population structure of L2 strains in endemic regions at a global scale, we compiled a collection of L2 whole-genome sequencing (WGS) data of 4425 clinical isolates from >50 studies representing 34 countries in Asia (86%), Eastern Europe (Russia and Belarus, 4%), Oceania (2%) and East/Southern Africa (8%). A summary of sources of isolates from previous studies is in Table S1. The full details are in Table S2, Supplementary Data Sheet 1. This dataset consisted of 3% L2.1 (n=150), 8% L2.2.2 (n=361), 88% L2.2.1 (n=3,913) and one unclassified L2 isolate, representing all of the twelve major L2 clades described so far [5, 26, 29] (Table S3). About 90% of the isolates were from Asia, about 76% of which were from China, Thailand or Vietnam. The isolates from these three countries represented diverse L2 strains rather than clonal or outbreak strains. A large proportion (26.4%) of the isolates could not be identified beyond Ancestral or Modern L2.2.1, 87% of which were Modern Beijing strains.

We identify >140,000 high-quality genome-wide SNPs (~3% of the genome), about 32% of which were phylogenetically informative. A maximum-likelihood (ML) phylogenetic tree

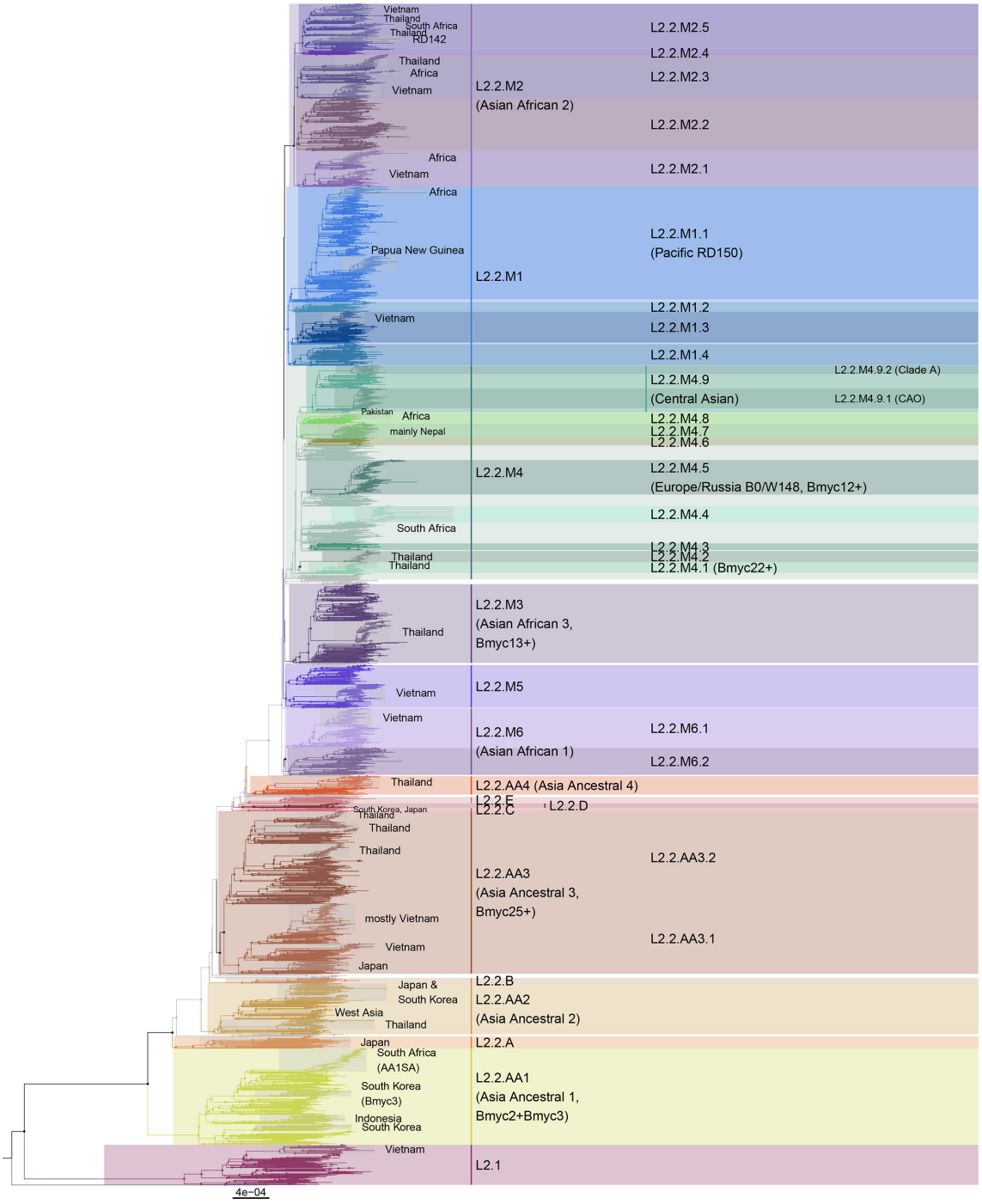

Fig. 1. Phylogenetic tree of 4,425 isolates in the discovery set estimated under the maximum-likelihood framework, rooted using the H37Rv reference strain (lineage 4). The first column of labels lists sixteen level-2 and level-3 genotypes (L2.2 not shown), the second column lists twenty-two level-4 genotypes and the last column lists two level-5 genotypes. Clades with notable features such as geographic specificity are highlighted. Previous or other names are given in parentheses; see Table 1 for references. Labels such as Bmyc13 +means the clade contains Bmyc13 [19] as a major subclade and non-Bmyc13 isolates at the base. Bmyc2 +Bmyc3 refers to a clade of both Bmyc2 and Bmyc3 strains. AA1SA refers to a South African subclade of L2.2.AA1 (32). RD142 refers to a subclade of L.2.2.M2.5 defined by the presence of a deletion in the RD142 region [14], which in turn contains Bmyc18 as a subclade. RD142 deletion was used to define L2.2.1.2 [17]. CAO, Central Asia Outbreak. An interactive version of the phylogeny is available online at Microreact (https://microreact.org/project/4P2iPeBx1Y66TyJfojNM1o).

Table 1. Revised list of phylogenetically informative genotypes of Mtb L2. Ancestral Beijing genotypes are listed as in the branching order in the phylogeny in Fig. 1. *Level* refers to the hierarchy level in the nomenclature; level 1 is the entire L2. *Other names* lists previously proposed genotypes based on phylogenetic analysis of WGS data. Genotypes with no other names are newly proposed in this study. Bmyc names are included to illustrate that most of the groups in this Bmyc scheme, an early SNP-based classification [19], do not correspond to monophyletic clades, except for those labelled in **bold** (also shown Fig. S1).

Genotype	Level	Other names	Bmyc name (Mestre)	Description
L2.1	2	Proto-Beijing [4]	**Proto-Bmyc1**	Non-Beijing L2
L2.2	2	Beijing		
L2.2.AA1	3	Asia Ancestral 1 [5], L2.2.2 (21), Bj-MG1 (3)	Bmyc2, **Bmyc3**	Contains clades from Japan/South Korea, Indonesia and a recent outbreak in South Africa (AA1SA clade)
L2.2.A	3	–	Bmyc4	Associated with Japan
L2.2.AA2	3	Asia Ancestral 2 [5]	Bmyc4	Contains large clades from Thailand and from Japan/South Korea
L2.2.B	3	–	Bmyc6	
L2.2.AA3	3	Asia Ancestral 3 [5], Bj-MG2 [4]	**Bmyc25**, Bmyc6	Contains several large clades from Thailand and from Vietnam
L2.2.AA3.1	4	–	**Bmyc25**	Vietnam-majority
L2.2.AA3.2	4	–	**Bmyc25**	Thailand-majority
L2.2.C	3	–	Bmyc26	Mostly from Japan and South Korea
L2.2.D	3	–	Bmyc26	Mostly from China
L2.2.E	3	–	Bmyc26	Mostly from China
L2.2.AA4	3	Asia Ancestral 4 [29], Bmyc26/10 [27]	Bmyc26/10	Mostly from Thailand
L2.2.M1	3	–		
L2.2.M1.1	4	Pacific RD150 [5], L2.2.1.1 [22]	Bmyc10	Contains large clades from Vietnam, Thailand, Papua New Guinea and South Africa
L2.2.M1.2	4	–	Bmyc10	Mostly from China and Vietnam
L2.2.M1.3	4	–	Bmyc10	Mostly from Vietnam
L2.2.M1.4	4	–	Bmyc10	Mostly from China
L2.2.M2	3	Asian African 2 [5]		
L2.2.M2.1	4	–	Bmyc10	Contains a large clade mostly from Vietnam, and a large clade from multiple African countries
L2.2.M2.2	4	–	Bmyc10	From diverse countries
L2.2.M2.3	4	–	Bmyc10	Contains large clades from Vietnam and from Thailand, and a small clade from South Africa and Mozambique
L2.2.M2.4	4	–	Bmyc10	A small clade, mostly from China
L2.2.M2.5	4	–	Bmyc10	Contains large clades from Vietnam and from Thailand, and a small clade of Bmyc18 within an RD142 clade (L2.2.1.2) [22]
L2.2.M3	3	Asian African 3 [26]	**Bmyc13**	Contains a large clade from Thailand that may be associated with drug resistance and recurring local outbreaks
L2.2.M4	3	–		
L2.2.M4.1	4	Bmyc22 [19, 29]	**Bmyc22**, Bmyc10	Mostly from Thailand
L2.2.M4.2	4	–	Bmyc10	All from Thailand
L2.2.M4.3	4	–	Bmyc10	

Continued

Table 1. Continued

Genotype	Level	Other names	Bmyc name (Mestre)	Description
L2.2.M4.4	4	–	Bmyc10	Mostly from South Africa
L2.2.M4.5	4	Europe/Russia B0/W148 [5], Clade B [12]	**Bmyc12**, Bmyc10	Mostly from Russia, Central Asia and Eastern Europe
L2.2.M4.6	4	–	Bmyc10	
L2.2.M4.7	4	–	Bmyc10	Mostly from Nepal
L2.2.M4.8	4	–	Bmyc10	Contains a clade from South Africa and Malawi
L2.2.M4.9	4	Central Asian [5]	Bmyc10	Mostly from Russia, Central Asia and Eastern Europe
L2.2.M4.9.1	5	Central Asia Outbreak (CAO) [5]	Bmyc10	
L2.2.M4.9.2	5	Clade A [12]	Bmyc10	
L2.2.M5	3	–	Bmyc10	
L2.2.M6	3	–		
L2.2.M6.1	4	Asian African 1 [5]	Bmyc10	
L2.2.M6.2	4	–	Bmyc10	

estimated from this SNP set recovers all the twelve previously described major L2 sublineages [26, 29] as monophyletic clades with 100% bootstrap support (Fig. 1). The phylogeny basally splits into two sublineages, L2.1 (proto-Beijing) and L2.2 (Beijing). We also identify a single isolate from South Korea that appears to be basal to the rest of the L2 isolates (**Supplementary Text**). It has a similar LSP profile as L2.1: an L2-specific deletion in RD105 and intact RD181 and RD207. It also has SNPs specific to L2 but does not have SNPs specific to L2.1 or any other L2 sublineages, consistent with its basal phylogenetic position. Overall, isolates from China, Thailand and Vietnam appear relatively evenly across the phylogeny and are present in most of the L2 sublineages (Table S3), suggesting long-term presence and a possible origin and diversification of L2 strains in the region.

L2.1 (proto-Beijing) mostly consists of isolates from China, Thailand and Vietnam, with a few isolates from Japan, Malaysia and Indonesia. There is no clear separation by country except for a distinct clade from Vietnam which could represent an emerging local strain. Most isolates have the typical SIT523 spoligotype (777777777777771).

Classification of L2.2 sublineages and the revised nomenclature

The remaining isolates belong to L2.2 (Beijing), which splits into two clades, L2.2.1 and L2.2.2, with over 90% of the isolates belonging to the former. The phylogeny of L2.2.1 has a cascading structure that ends with a large star-shaped clade of the Modern Beijing strains. L2.2 isolates that do not belong to the Modern Beijing genotype are sometimes referred to as Ancestral Beijing strains, which are not monophyletic.

The cascading structure of L2.2.1 phylogeny poses a challenge in defining genotypes that both reflect their hierarchical phylogenetic relationships and capture potentially important circulating strains without having too many nested levels of hierarchy in the nomenclature, which would preclude wide adoption of the scheme. We make a compromise by discarding the L2.2.1 and L2.2.2 nomenclature, similar to Rutaihwa *et al*. [30] and proposing fifteen level-3 genotypes of L2.2 (Table 1). The criteria are described in Methods. Briefly, a genotype must correspond to a monophyletic clade with bootstrap support value at least 90%, and must uniquely share at least one SNP (relative to the reference H37Rv strain). Our scheme captures >98% of L2.2 isolates. The remaining <2% are left as unclassified Ancestral or Modern Beijing strains due to either insufficient phylogenetic structure or the number of isolates is too small to confidently identify clade-specific SNPs. We note that there are a few well-supported clades of Ancestral Beijing strains with too few isolates (<10) that could emerge as distinct genotypes when more genomes are added.

We next describe the nomenclature for L2.2 sublineages. Each genotype has a combination of capital letters and numbers as a label for the level-3 classification. If a genotype has been previously described, multiple letters and numbers are used in a meaningful way, for instance, to reflect the existing name. For example, we use L2.2.AA1 for Asia Ancestral 1 [5]. For new Ancestral Beijing genotypes, we use a single letter starting from A (e.g. L2.2.A, L2.2.B). For Modern Beijing genotypes, we use the letter M followed by a number (e.g. L2.2.M1, L2.2.M2). Further classification of each genotype can be added at the fourth or fifth level if appropriate (e.g. L2.2.AA3.1, L2.2.M1.1). This nomenclature system is designed to be simple and human-readable, compatible with existing genotypes, and extendible when new genotypes are discovered.

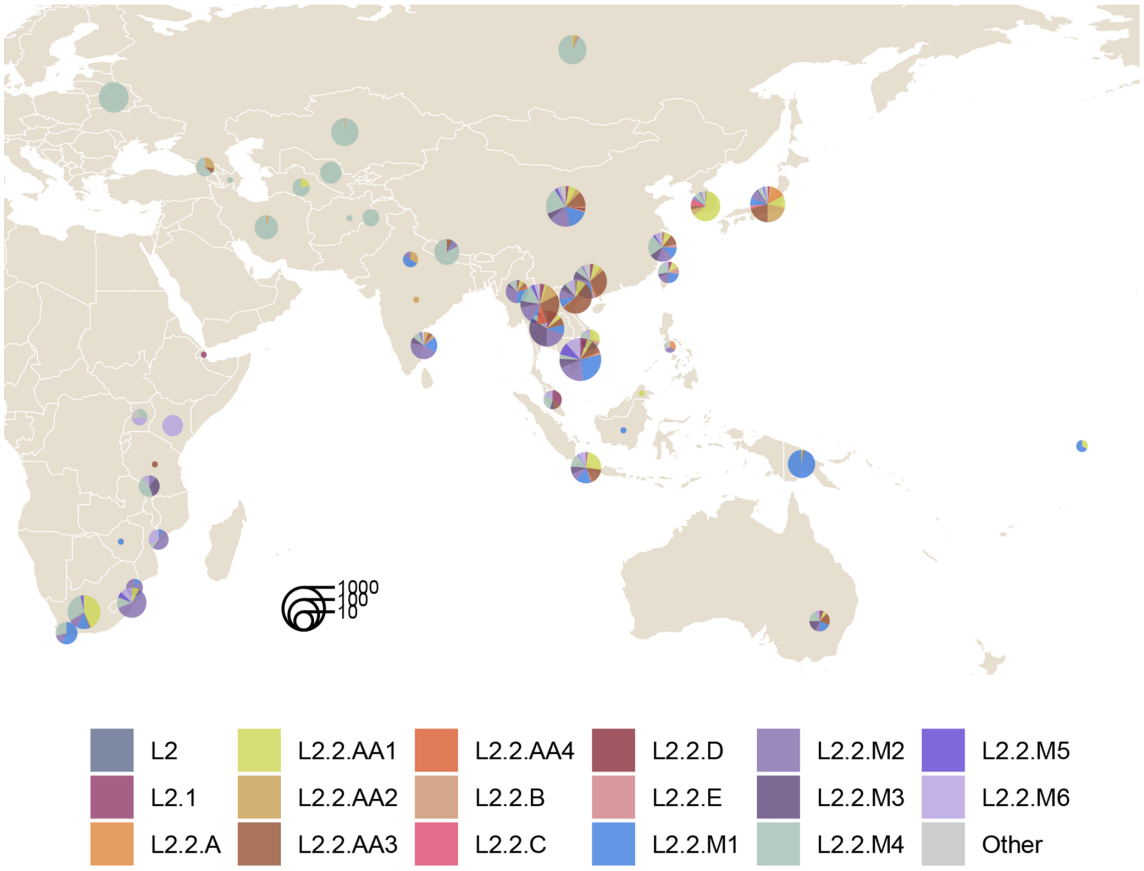

Fig. 2. Geographical distribution Mtb L2 by country and region. Pie charts show proportions of isolates from each location by sublineages. Pie sizes are proportional to the total number of isolates from each location. An interactive version of this map is available online at Microreact (https://microreact.org/project/4P2iPeBx1Y66TyJfojNM1o).

Revised Ancestral Beijing genotypes

L2.2 is classified into nine Ancestral Beijing genotypes and six modern Beijing genotypes (Table 1). We propose the following names for the nine genotypes of Ancestral Beijing strains: L2.2.AA1, L2.2.A, L2.2.AA2, L2.2.B, L2.2.AA3, L2.2.C, L2.2.D, L2.2.E and L2.2.AA4 (in their cascading branching order in the phylogeny in Fig. 1). L2.2.AA1 coincides with L2.2.2. The remaining eight clades share the RD181 deletion and belong to L2.2.1 in the scheme of Coll *et al.* [22].

L2.2.AA1 (Asia Ancestral 1) contains several distinct clades associated with specific countries (Table S3, Fig. S2), such as clades associated with South Korea/Japan, Indonesia and South Africa. We find one Russian isolate (SRR6256978) from Buryatia in eastern Siberia nested within a clade consisted entirely of isolates from South Korea and Japan, suggesting that L2.2.AA1 could have originated in East Asian and spread to North Asia [47]. Moreover, L2.2.AA1 appears to be the most common strain (64%) among South Korean isolates. One South Korean-majority clade coincides with the Bmyc3 genotype [19], nested within Bmyc2. Bmyc3 has been associated with an outbreak strain in South Korea known as the K strain [26, 48]. There is a large South African clade with a distinct cascading structure, short terminal branches and poor internal branch supports, indicative of recent local outbreaks. This clade, referred to as AA1SA, is also associated with highly drug-resistant strains that have become endemic to South Africa [35]. Further classification of AA1SA has been described [35] but we did not observe any clear phylogenetic structure of subclades of AA1SA in our dataset.

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.

L2.2.AA2 (Asia Ancestral 2) contains isolates mostly from China, Thailand and Japan, with two large clades, Thailand-majority and Japan-majority. There is a small clade (n=9) consisting entirely of isolates from western Asian countries (Georgia, Kazakhstan, Iran) and Russia, with a long supporting branch. This clade is a part of a larger clade that has been strongly associated with drug resistance (MDR and XDR) in Russia [47, 49] (Fig. S3).

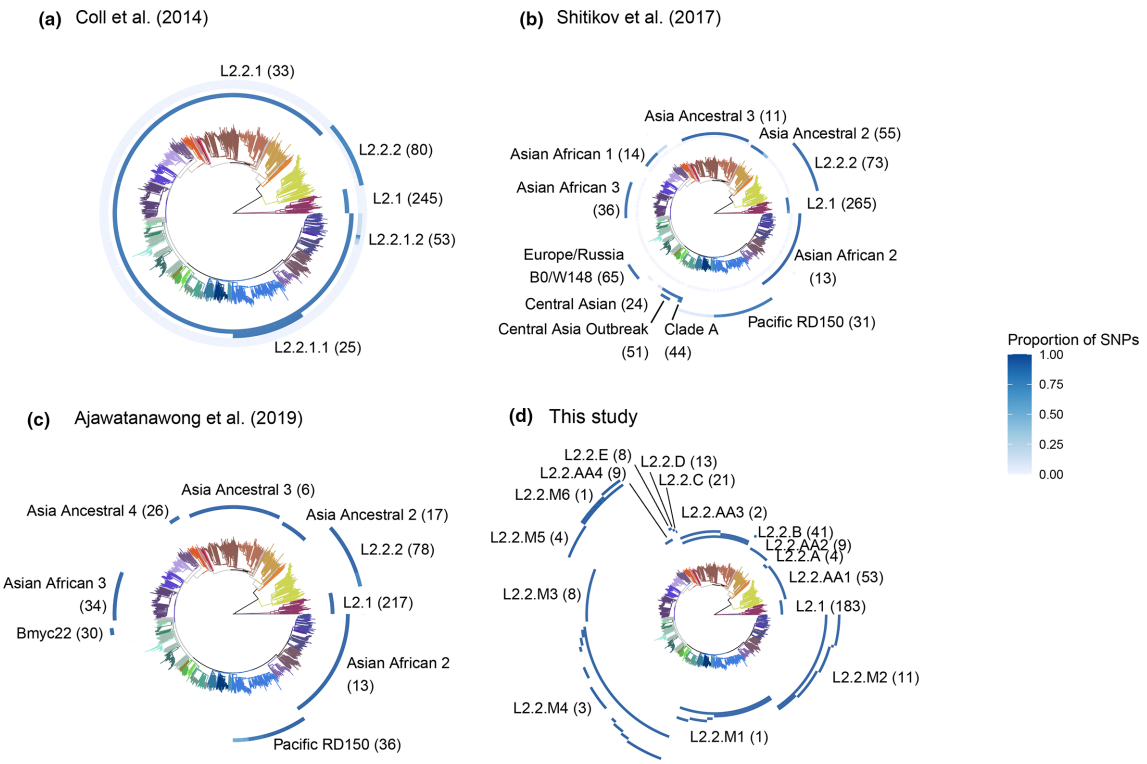

Fig. 3. Comparison with three existing schemes for SNP-based genotyping of 4,425 L2 strains. Colour intensity of the arcs represents the proportion of lineage-specific SNPs in each scheme present in our genomic dataset. It is always one in our scheme (d) since the same dataset was used to derive the scheme. They are mostly close to one in other schemes, except for a few places, e.g. some 2.2.M1.1 in (c) and 2.2.1.2 in (a). The number of lineage-specific SNPs are given in parentheses after each genotype name. For (d), only level-3 genotypes are labelled. Notice several genotypes in the previous schemes are defined by SNPs which are not actually specific to the intended genotype because the samples used were not sufficiently representative of the actual strain diversity, for example, L2.2.1.2 in Coll *et al.* [22] or Pacific RD150 in Shitikov *et al.* [26].

The rest of the phylogeny is the sister clade of L2.2.AA2 where isolates share the variant CGG/GGG at *mutT4* codon 48 (4393590 C/G) [19]. L2.2.B is a small Bmyc6 clade with a unique SIT250 spoligotype (000000000000371). There are two other distinct Bmyc6 clades left unclassified Ancestral 2.2 (Beijing) since the numbers of isolates are too small (<10).

L2.2.AA3 (Asia Ancestral 3) is the most common genotype among the Ancestral Beijing strains. It contains two major clades, designated as L2.2.AA3.1 (Vietnam-majority) and L2.2.AA3.2 (Thailand-majority), and a small basal clade of only two isolates from Thailand. The two major clades are Bmyc25, characterized by the variant 1477522 C/A at *ogt* codon 37, while the basal isolates are Bmyc6 [26, 29]. Both clades have smaller but comparable proportions of isolates from China (25%) and Japan (10%).

We assign three small Bmyc26 clades L2.2.C, L2.2.D and L2.2.E. L2.2.C is associated with Japan and South Korea while L2.2.D and L2.2.E consist of mostly Chinese isolates. A few Bmyc26 isolates have no strong phylogenetic structure and are left as unclassified Ancestral 2.2 (Beijing).

L2.2.AA4 (Asia Ancestral 4) represents the final major clade of Ancestral Beijing strains before transitioning to Modern Beijing strains. It is dominated by Thai isolates (~70%), with small proportions of isolates from China and Vietnam. It is probably associated with ethnic minority groups who distribute in the three countries [29]. L2.2.AA4 has the nonsynonymous variant GGA/CGA at *mutT2* codon 58 (1286766 G/C) previously thought to be specific to Modern Beijing strains but lack the synonymous variant GGG/GGA at *ogt* codon 12 (1477596 C/T) shared by all Modern Beijing isolates.

Six major clades of Modern Beijing genotypes

The phylogenetic tree of Modern Beijing strains is star-like, suggestive of recent population expansions. Seven major genotypes have been described previously: Asian African 1–3, Bmyc22, Pacific RD150, Central Asian, and Europe/Russia B0/W148 [5, 26, 30]. Our results suggested that most of the Modern Beijing strains can be clustered into six well-supported clades, labelled L2.2.M1–L2.2.M6, where 'M' designates 'Modern Beijing'. All previously described genotypes fall into five of these clades while L2.2.M5 does

not appear to contain any known genotype (Table 1). This six-group scheme is able to identify ~99% of the Modern Beijing isolates (L2.2.M1 23%, L2.2.M2 24%, L2.2.M3 10%, L2.2.M4 27.5%, L2.2.M5 5.5% and L2.2.M6 9%), with ~1% of the isolates remain as unclassified Modern Beijing strains. Four clades, L2.2.M1, L2.2.M2, L2.2.M4 and L2.2.M6, are further classified into subclades at the fourth level. L2.2.M1 and L2.2.M4 have some isolates left unassigned at the fourth level due to poor phylogenetic resolution (18/670=3% of L2.2.M1 and 199/794=25% of L2.2.M4).

L2.2.M1 contains four subclades: L2.2.M1.1–L2.2.M1.4. L2.2.M1.1 coincides with the Pacific RD150 genotype [5], also called L2.2.1.1 [22], characterized by the deletion of RD150. It is the main subclade of L2.2.M1 (422/672=63%) and is present in several countries around the Indian Ocean and in the Pacific Ocean (Table S3, Fig. S2). Its geographical distribution is more widespread than previously thought [5, 26], with high prevalence in East Asia and mainland Southeast Asia. Basal L2.2.M1.1 isolates tend to come from China whereas more terminal isolates are mainly from Thailand and Vietnam. Two notable large subclades could be associated with local outbreaks, one in Papua New Guinea and another in South Africa. The remaining three subclades, L2.2.M1.2–L2.2.M1.4, are dominated by isolates from China and Vietnam. L2.2.M1.3 also contains a large Vietnamese clade.

L2.2.M2 is Asian African 2 [5]. It has a distinct long supporting branch compared with other M clades and clear within-clade separation into five subclades, designated as L2.2.M2.1–L2.2.M2.5. L2.2.M2.4 is a small clade of mostly Chinese isolates. The other four clades have substantial proportions of isolates from China, Vietnam, Thailand and South Africa, and contain large clades associated with each country. L2.2.M2.5 includes the so-called L2.2.1.2 as a small clade (25/146=14% of L2.2.M2.5), defined based on the shared RD142 deletion (~2.85 kb) [22]. Bmyc18 [19] forms a small clade within RD142 (18 out of 25), and is thus not synonymous with RD142 as previously suggested [26]. We do not assign a specific genotype for this group since it only constitutes a small fraction of L2.2.M2.5, with surprisingly little geographic specificity.

L2.2.M3 is Asian African 3 [26]. A few basal isolates are Bmyc10 while the rest of the clade are Bmyc13 [26, 29]. L2.2.M3 contains three major clades, each of which is associated with a specific country: China, Vietnam and Thailand. The Thai clade consists exclusively of samples from Thailand forming a large genomic cluster that could represent recent MDR outbreaks in Thailand [50]. All African isolates in L2.2.M3 come from Malawi in our dataset.

L2.2.M4 is the largest and the most diverse Modern clade, with several outbreak strains associated with distinct geography. We further assign nine genotypes: L2.2.M4.1–L2.2.M4.9, three of which have been described in the literature: L2.2.M4.1 (Bmyc22 [19, 29]), L2.2.M4.5 (Europe/Russia B0/W148 [4, 5, 51]) and L2.2.M4.9 (Central Asian [4, 5]). These clades are extended to include more basal isolates when possible. The subclades of L2.2.M4 tend to show high geographic specificity and short terminal branches. For example, L2.2.M4.5 and L2.2.M4.9 are the only two L2 genotypes with high prevalence in Central Asia, Russia and Eastern Europe but are uncommon elsewhere, whereas L2.2.M4.4 and L2.2.M4.8 are mainly associated with Africa. 25% (199/794) of L2.2.M4 isolates, mostly from China, are still left as unclassified.

L2.2.M4.5 consists of two clades: Europe/Russia B0/W148 [4, 5, 51], which is Bmyc12 (Fig. S1), and a small Bmyc10 clade of isolates from China. Other synonymous names of Europe/Russia B0/W148 include Clade B [12], East European 2 [4] and Clonal Complex 2 (CC2) [5]. It likely originated in Siberia in North Asia and spread across Central Asia, Russia and countries associated with the former Soviet Union [52]. L2.2.M4.9 is the Central Asian clade or Clonal Complex 1 (CC1) [5], also known as East European 1 [4], Central Asian/Russian and 94–32 cluster [53, 54]. Similar to L2.2.M4.5, it is geographically restricted to Central Asia, Russia and Eastern Europe, but with higher prevalence in Central Asia compared with L2.2.M4.5. It contains two major subclades: (i) L2.2.M4.9.1, known as the Central Asia Outbreak (CAO), which is more prevalent in Central Asian countries (such as Kazakhstan and Uzbekistan) [55], and (ii) L2.2.M4.9.2, known as Clade A, which is more prevalent in Russia and Eastern Europe [12] (Fig. S2).

The remaining L2.2.M4 subclades are smaller and more geographically restricted. L2.2.M4.1 and L2.2.M4.2 are sister clades. L2.2.M4.1 mainly comprises Bmyc22 (37/41=90%) and a few basal Bmyc10 isolates. Most isolates are from Thailand which associated with HIV infection [29]. L2.2.M4.2 is specific to Thailand and could represent a local outbreak cluster. L2.2.M4.3, L2.2.M4.6 and L2.2.M4.8 are dominated by isolates from China and tend to have long terminal branches, suggesting their long-term presence in East Asia. L2.2.M4.6 contains a Vietnamese clade while L2.2.M4.8 contains a large clade associated with several African countries including South Africa, Mozambique and Malawi. L2.2.M4.4 is a large South African clade, with one basal isolate from Myanmar, suggesting a possible introduction from Asia. L2.2.M4.7 is Nepal-majority (34/58=59%) and contains a small Thai subclade.

Finally, L2.2.M5 and L2.2.M6 are relatively small Modern clades with isolates mostly from Vietnam, Thailand and China. Both also contain several African clades nested among Asian isolates. L2.2.M6 consists of two subclades, L2.2.M6.1 and L2.2.M6.2. L2.2.M6.1 is previously described as Asian African 1 [5].

Population diversity and divergence

To characterize the genetic diversity within and between the proposed genotypes, we calculated several summary statistics for each genotype and for pairs of genotypes. For most genotypes, the nucleotide diversity is 0.003% on average (Fig. S4), compared with ~0.01% for main Mtb lineages (L1–L7) [20]. Early branching genotypes such as L2.1, L2.2.AA1 and L2.2.A have the highest nucleotide

diversity, ~0.006%, while L2.2.M4.1 (Bmyc22), L2.2.M4.2 (Thailand), L2.2.M4.4 (South Africa) and L2.2.M4.5 (Europe/Russia B0/W148) subclades of L2.2.M4 have the lowest diversity of about 0.001% or lower. Pairwise SNP distances between isolates, averaged within and across genotypes, indicate a clear distinction between four groups of genotypes: L2.1, early Ancestral Beijing genotypes (L2.2.AA1 and L2.2.A), remaining Ancestral Beijing genotypes and Modern Beijing genotypes (Fig. S5a–c, Table S4). The median pairwise SNP distances between genotypes for these four groups are as follows: ~800–900 SNPs between L2.1 and L2.2, ~400–500 SNPs for between the early Ancestral Beijing genotypes and the other L2.2, 300 SNPs between the rest of Ancestral Beijing and Modern Beijing genotypes, and ~200 SNPs among the Modern Beijing genotypes. Within genotypes, the pairwise SNP distances are around 100–200 on average.

We further investigate genetic differentiation between genotypes from genome-wide average of the fixation index (F_{ST}) between genotypes and principal components analysis (PCA) of 4,425 isolates. We find relatively high F_{ST} across genotypes, with L2.1 being most differentiated from other genotypes (F_{ST} ~0.7) (Fig. S6a). The Modern Beijing genotypes are more genetically similar to each other (F_{ST} ~0.2). Exceptions are clades with a long supporting branch and short terminal branches such as L2.2.M4.2, L2.2.M4.4 and L2.2.M4.5 (see Fig. 1), which are more differentiated from other Modern Beijing genotypes (F_{ST} ~0.5–0.6). Ancestral Beijing genotypes tend to have intermediate F_{ST} values of ~0.4–0.5. PCA confirms strong clustering of isolates by genotype, particularly L2.1 and Ancestral Beijing genotypes (Fig. S6b), consistent with the observed patterns from the phylogeny (Fig. 1) and other summary statistics (Fig. S5). PCA on the Modern Beijing strains supports clustering of the six genotypes L2.2.M1–L2.2.M6 (Fig. S6c).

Geographical distribution of Mtb L2 genotypes

We observe geographic specificity among Mtb L2 strains (Figs 2 and S2). Isolates from China, Thailand and Vietnam tend to be basal in the phylogeny, suggesting the long-term presence of diverse L2 strains in the region. Japan and South Korea have higher prevalence of Ancestral Beijing strains, especially L2.2.A and L2.2.C. L2.2.M4 clades show strongest geographical clustering, including two well-known Europe/Russia/Central Asia-majority clades L2.2.M4.5 (Europe/Russia B0/W145) and L2.2.M4.9 (Central Asian), two Africa-majority clades (L2.2.M4.4 and L2.2.M4.8), a Nepal-majority clade (L2.2.M4.7) and two Thailand-majority clades (L2.2.M4.1 and L2.2.M4.2). African isolates were present in all six Modern Beijing clades (L2.2.M1–L2.2.M6), and L2.2.AA1. They always appear as tip clades, suggesting multiple introductions of L2 strains from Asia into Africa, most likely via maritime trade routes [25, 30]. We caution that non-systematic sampling of the isolates can bias inference about the geographic composition of genotypes.

Genomic clusters across L2 phylogeny

To gain insights into possible transmission and outbreak clusters of Mtb L2, we identify genomic clusters on the phylogeny as clades that include isolates linked by at most 12 SNP differences [40, 41]. We find that 1065 isolates grouped into 250 clusters, with median cluster size of two isolates (Table S5). Only thirteen clusters have at least ten isolates (Fig. S7). Each cluster was always from a single country. There is no clear association between genotype and cluster size or the number of cluster. The six largest clusters (>30 isolates) are belong to different genotypes in both Ancestral and Modern Beijing groups (L2.2.AA1, L2.2.AA3.2, L2.2.M3, L2.2.M4.4, L2.2.M2.3, and L2.2.M4.2) and are from two countries: Thailand and South Africa (Table S5, Fig. S8).

Revised SNP-based genotyping scheme for Mtb L2

We identify SNPs specific to each genotype of Mtb L2 at three hierarchical levels when applicable (Table 1). The numbers of genotype-specific SNPs are in Table S6. The full list is in Table S7. We also select a few SNPs as genotyping (or barcoding) SNPs for each genotype based on properties of the variants that tend to be evolutionarily stable as well as consistency with existing schemes. In particular, we prefer synonymous SNPs in T cell epitopes, in essential genes or in the third codon position. These genotyping SNPs are indicated in the column 'barcoding' in Table S7. The barcoding SNPs for each genotype were summarized in Table S8.

Validation of the genotyping scheme

We validated the proposed genotyping scheme (Table S7) on an independent test set of 1,207 isolates as, representing 32 countries from Asia, Africa, Europe and Americas (Table S1). The genotype identity was confirmed via phylogenetic tree inference of isolates from both discovery and test sets (Fig. S9). All genotypes are recovered as well-supported clades in the phylogeny of the combined dataset, demonstrating the stability of our genotypes. From the test set, we identify most of the genotypes in our proposed scheme, the majority (87%) of which are Modern Beijing strains (Table S9). Only six isolates (0.5%), all from Asia, are unclassified L2.2 strains in our scheme, in contrast with 180 isolates (15%) based on existing schemes. Genotypes not present in the test set are L2.2.B, L2.2.D, L2.2.M4.4 and L2.2.M4.6. These clades are small or specific to regions not included in the test set.

DISCUSSION

Here, we provide a revised hierarchical genotyping scheme for Mtb L2 based on phylogenetic and population genetic analyses of a collection of isolates representative of the endemic areas in Asia, Eastern Europe, Africa and Oceania. This collection is significantly more diverse than the previous works. Our nomenclature system is designed to capture the hierarchical relationships between genotypes while being human-readable as well as being backward compatible with genotypes defined based on genetic and epidemiological evidence. It classifies L2 into two genotypes at the second level of the hierarchy, 15

genotypes at the third level, 22 genotypes at the fourth level and two genotypes at the fifth level (Table 1). This represents a substantial improvement over existing genotyping schemes in terms of coverage: over 98% of the studied L2 instead of just under 75% in the discovery set (n=4,425) (Fig. 3), and 99.5% versus 85% in an independent test set of globally diverse strains (n=1,207).

In order to facilitate rapid identification of L2 sublineages, we summarize the proposed barcoding SNPs in Table S8. Although it is not always necessary, barcoding of every level of classification should be identified. As new variants of *M. tuberculosis* are likely to emerge, confirmation of the barcoding results with the whole set of sublineage specific SNPs should be considered, especially for isolates from the countries that WGS data are still scarce.

Since the Beijing strains are possibly expanding, as exemplified by several country-associated clades, it is possible that more widespread applications of WGS will lead to discovery of new large clades. The genomic framework presented here can be updated and extended to include new genotypes as they emerge or more data are collected. The finding that L2.2.M4 encompasses isolates from highly diverse regions, including several well-known genotypes associated with outbreaks or drug resistance that are distributed over all the studied areas, suggests intriguing properties of the L2.2.M4 genotypes that deserve further investigation.

One major limitation of our work is the sampling bias. We relied mainly on publically available WGS data and previously published reports, and the majority of isolates were from China, Vietnam and Thailand. Non-systematic sampling across geographic regions can bias inference about geographic composition (Fig. S2) and spread of genotypes. More WGS data from under-represented countries and regions such as Myanmar, Laos, Cambodia, Indonesia, many African countries, as well as South Asian countries where TB incidence is highest, may uncover more phylogenetic structure and new genotypes. As another limitation, we did not investigate genotype-specific insertions or deletions. Some large deletions have traditionally been used to classify L2 strains into a few groups [14, 22], but there could be more. Finally, our genotyping scheme could be refined and improved with more data, including WGS, epidemiological and clinical data, to support the significance of genotypes beyond population and phylogenetic structure. For instance, there are clear distinct subclades within several genotypes including L2.2.AA3, L2.2.M3 and L2.2.M4.9, some of which could potentially be associated with local outbreaks (Fig. S7) or multidrug resistance. In this case, new genotypes could be added by continuing the current nomenclature, e.g. L2.2.AA3.1.1, L2.2.M3.1 or L2.2.M4.9.3.

With increasing routine use of whole-genome sequencing of Mtb, our genotyping and nomenclature framework provides a solution to genotyping of globally diverse Mtb L2 strains. Having such a fine-scale genotyping scheme opens up opportunities for enriched comparisons between studies, improving surveillance of transmission and possibly detecting the spread of drug resistant strains. This can also lead to better understanding of local strain diversity in a global framework, more meaningful interpretations of WGS data in clinical and public health settings, better genotype–phenotype associations, better design of drugs and vaccines, as well as improvements in designing and implementing TB control strategies tailored to local contexts. Detailed genotyping can also be useful for further genomic investigation, such as homoplastic mutation identification [56].

Funding information
This work is supported by Mahidol University-Multidisciplinary Research Center grant, the Emerging Infectious Disease program, National Science and Technology Development Agency, Department of Medical Sciences, Ministry of Public Health and Fogarty International Center, US-National Institutes of Health. (Grant number D43TW009522).

Author contributions
Y.T. and P.P. conceived the study. S.M., H.Y., H.M.W.M., Z.C. and V.C. provided sequence data. Y.T. curated the data, performed the formal analysis and visualized the results. Y.T. and P.P. wrote the original draft. All authors reviewed the paper.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Thawornwattana Y, Mahasirimongkol S, Yanai H, Maung H, Cui Z, *et al*. Revised nomenclature and SNP barcode for Mycobacterium-tuberculosis lineage 2. *Figshare* 2021.
2. Wiens KE, Woyczynski LP, Ledesma JR, Ross JM, Zenteno-Cuevas R, *et al*. Global variation in bacterial strains that cause tuberculosis disease: A systematic review and meta-analysis. *BMC Med* 2018;16:196.
3. O'Neill MB, Shockey A, Zarley A, Aylward W, Eldholm V, *et al*. Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. *Mol Ecol* 2019;28:3241–3256.
4. Luo T, Comas I, Luo D, Lu B, Wu J, *et al*. Southern East Asian origin and coexpansion of *Mycobacterium tuberculosis* Beijing family with Han Chinese. *Proc Natl Acad Sci U S A* 2015;112:8136–8141.
5. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, *et al*. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat Genet* 2015;47:242–249.
6. Cohen KA, Manson AL, Abeel T, Desjardins CA, Chapman SB, *et al*. Extensive global movement of multidrug-resistant *M. tuberculosis* strains revealed by whole-genome analysis. *Thorax* 2019;74:882–889.
7. Ribeiro SCM, Gomes LL, Amaral EP, Andrade MRM, Almeida FM, *et al*. *Mycobacterium tuberculosis* strains of the modern sublineage of the beijing family are more likely to display increased virulence than strains of the ancient sublineage. *J Clin Microbiol* 2014;52:2615–2624.
8. Rajwani R, Yam WC, Zhang Y, Kang Y, Wong BKC, *et al*. Comparative whole-genomic analysis of an ancient L2 lineage *Mycobacterium tuberculosis* reveals a novel phylogenetic clade and common genetic determinants of hypervirulent strains. *Front Cell Infect Microbiol* 2017;7:539.
9. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, *et al*. Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat Genet* 2018;50:849–856.
10. Liu Q, Ma A, Wei L, Pang Y, Wu B, *et al*. China's tuberculosis epidemic stems from historical expansion of four strains of *Mycobacterium tuberculosis*. *Nat Ecol Evol* 2018;2:1982–1992.
11. Fine PEM, Crampin AC, Houben R, Mzembe T, Mallard K, *et al*. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife* 2015;2015:e05166.

12. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, *et al*. Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 2014;46:279–286.
13. Eldholm V, Pettersson JHO, Brynildsrud OB, Kitchen A, Rasmussen EM, *et al*. Armed conflict and population displacement as drivers of the evolution and dispersal of *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* 2016;113:13881–13886.
14. Tsolaki AG, Gagneux S, Pym AS, Goguet De La Salmoniere YOL, Kreiswirth BN, *et al*. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of *Mycobacterium tuberculosis*. *J Clin Microbiol* 2005;43:3185–3191.
15. Smittipat N, Billamas P, Palittapongarnpim M, Thong-On A, Temu MM, *et al*. Polymorphism of variable-number tandem repeats at multiple loci in *Mycobacterium tuberculosis*. *J Clin Microbiol* 2005;43:5034–5043.
16. 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.
17. Ebrahimi-Rad M, Bifani P, Martin C, Kremer K, Samper S, *et al*. Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerg Infect Dis* 2003;9:838–845.
18. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, *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. *J Bacteriol* 2006;188:759–772.
19. Mestre O, Luo T, Dos Vultos T, Kremer K, Murray 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:e16020.
20. Coscolla M, Gagneux S. Consequences of genomic diversity in Mmycobacterium tuberculosis. *Semin Immunol* 2014;26:431–444.
21. Zhang H, Li D, Zhao L, Fleming J, Lin N, *et al*. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat Genet* 2013;45:1255–1260.
22. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, *et al*. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* 2014;5:4812.
23. Mokrousov I, Narvskaya O, Otten T, Vyazovaya A, Limeschenko E, *et al*. Phylogenetic reconstruction within *Mycobacterium tuberculosis* Beijing genotype in northwestern Russia. *Res Microbiol* 2002;153:629–637.
24. Plikaytis BB, Marden JL, Crawford JT, Woodley CL, Butler WR, *et al*. Multiplex PCR assay specific for the multidrug-resistant strain W of *Mycobacterium tuberculosis*. *J Clin Microbiol* 1994;32:1542-1546
25. Mokrousov I, Ho ML, Otten T, Nguyen NL, Vyshnevskyi B, *et al*. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: Clues from human phylogeography. *Genome Res* 2005;15:1357–1364.
26. Shitikov E, Kolchenko S, Mokrousov I, Bespyatykh J, Ischenko D, *et al*. Evolutionary pathway analysis and unified classification of East Asian lineage of *Mycobacterium tuberculosis*. *Sci Rep* 2017;7:9227.
27. Liu Q, Luo T, Dong X, Sun G, Liu Z, *et al*. Genetic features of *Mycobacterium tuberculosis* modern Beijing sublineage. *Emerg Microbes Infect* 2016;5:e14.
28. Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, *et al*. Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med* 2020;12:114.
29. Ajawatanawong P, Yanai H, Smittipat N, Disratthakit A, Yamada N, *et al*. A novel Ancestral Beijing sublineage of *Mycobacterium tuberculosis* suggests the transition site to Modern Beijing sublineages. *Sci Rep* 2019;9:13718.
30. Rutaihwa LK, Menardo F, Stucki D, Gygli SM, Ley SD, *et al*. Multiple introductions of *Mycobacterium tuberculosis* Lineage 2-Beijing into Africa over centuries. *Front Ecol Evol* 2019;7:112.
31. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
32. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ARXIV. 2013. http://arxiv.org/abs/1303.3997
33. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, *et al*. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. [Preprint] 2018.
34. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 2012;6:80–92.
35. Klopper M, Heupink TH, Hill-Cawthorne G, Streicher EM, Dippenaar A, *et al*. A landscape of genomic alterations at the root of a near-untreatable tuberculosis epidemic. *BMC Med* 2020;18:24.
36. Xia E, Teo YY, Ong RTH. SpoTyping: Fast and accurate *in silico Mycobacterium* spoligotyping from sequence reads. *Genome Med* 2016;8:19.
37. Page A, Alikhan N-F, Strinden M, Le Viet T, Skvortsov T. Rapid *Mycobacterium tuberculosis* spoligotyping from uncorrected long reads using Galru. *bioRxiv* 2020.
38. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, *et al*. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;37:1530–1534.
39. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017;14:587–589.
40. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, *et al*. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study. *Lancet Infect Dis* 2013;13:137–146.
41. Yang C, Luo T, Shen X, Wu J, Gan M, *et al*. Transmission of multidrug-resistant *Mycobacterium tuberculosis* in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *Lancet Infect Dis* 2017;17:275–284.
42. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: The impact of rare variants. *Genome Res* 2013;23:1514–1521.
43. Miles A, Harding NJ. Scikit-allel. *Zenodo* 2017.
44. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, *et al*. Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. *mBio* 2017;8:e02133-16.
45. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, *et al*. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* 2010;42:498–503.
46. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, *et al*. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019;47:43.
47. Mokrousov I, Sinkov V, Vyazovaya A, Pasechnik O, Solovieva N, *et al*. Genomic signatures of drug resistance in highly resistant Mycobacterium tuberculosis strains of the early ancient sublineage of Beijing genotype in Russia. *Int J Antimicrob Agents* 2020;56:106036.
48. Han SJ, Song T, Cho Y-J, Kim J-S, Choi SY, *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.
49. Mokrousov I, Vyazovaya A, Pasechnik O, Gerasimova A, Dymova M, *et al*. Early ancient sublineages of Mycobacterium tuberculosis Beijing genotype: unexpected clues from phylogenomics of the pathogen and human history. *Clin Microbiol Infect* 2019;25:1039.
50. Regmi SM, Chaiprasert A, Kulawonganunchai S, Tongsima S, Coker OO, *et al*. Whole genome sequence analysis of multidrug-resistant *Mycobacterium tuberculosis* Beijing isolates from an outbreak in Thailand. *Mol Genet Genomics* 2015;290:1933–1941.

51. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. *Trends Microbiol* 2002;10:45–52.
52. Mokrousov I. Insights into the origin, emergence, and current spread of a successful Russian clone of *Mycobacterium tuberculosis*. *Clin Microbiol Rev* 2013;26:342–360.
53. Shitikov E, Vyazovaya A, Malakhova M, Guliaev A, Bespyatykh J, *et al*. Simple assay for detection of the Central Asia outbreak clade of the *Mycobacterium tuberculosis* Beijing genotype. *J Clin Microbiol* 2019;57:e00215-19.
54. Mokrousov I, Chernyaeva E, Vyazovaya A, Skiba Y, Solovieva N, *et al*. Rapid assay for detection of the epidemiologically important Central Asian/Russian strain of the *Mycobacterium tuberculosis* Beijing genotype. *J Clin Microbiol* 2018;56:e01551-17.
55. Merker M, Barbier M, Cox H, Rasigade JP, Feuerriegel S, *et al*. Compensatory evolution drives multidrug-resistant tuberculosis in central Asia. *Elife* 2018;7:e38200.
56. Tantivitayakul P, Ruangchai W, Juthayothin T, Smittipat N, Disratthakit A, *et al*. Homoplastic single nucleotide polymorphisms contributed to phenotypic diversity in *Mycobacterium tuberculosis*. *Sci Rep* 2020;10:8024.