Whole-genome single nucleotide variant phylogenetic analysis of *Mycobacterium tuberculosis* Lineage 1 in endemic regions of Asia and Africa

Thidarat Netiku1,2, Yuttapong Thawornwattana2,3, Surakameth Mahasirimongkol4, Hideki Yanai5, Htet Myat Win Maung6,7, Virasakdi Chongsuvivatwong7 & Prasit Palittapongarnpim2,8*

*Mycobacterium tuberculosis* (*Mtb*) lineage 1 (L1) contributes considerably to the disease morbidity. While whole genome sequencing (WGS) is increasingly used for studying *Mtb*, our understanding of genetic diversity of L1 remains limited. Using phylogenetic analysis of WGS data from endemic range in Asia and Africa, we provide an improved genotyping scheme for L1. Mapping deletion patterns of the 68 direct variable repeats (DVRs) in the CRISPR region of the genome onto the phylogeny provided supporting evidence that the CRISPR region evolves primarily by deletion, and hinted at a possible Southeast Asian origin of L1. Both phylogeny and DVR patterns clarified some relationships between different spoligotypes, and highlighted the limited resolution of spoligotyping. We identified a diverse repertoire of drug resistance mutations. Altogether, this study demonstrates the usefulness of WGS data for understanding the genetic diversity of L1, with implications for public health surveillance and TB control. It also highlights the need for more WGS studies in high-burden but underexplored regions.

*Mycobacterium tuberculosis* (*Mtb*) is classified into 9 lineages1–3, with lineages 1–4 (L1–L4) being widely distributed around the world. *Mtb* L1, or the India-Oceanic lineage, is the only *TbD1*-positive *Mtb* lineage with high incidences in many high-burden countries around the Indian Ocean4. Although whole genome sequencing (WGS) is increasingly used for studying *Mtb*, our understanding of L1 genetic diversity and epidemiology in high-burden settings remains limited.

*Mtb* L1 was first classified into two main groups based on whole-genome SNPs, denoted L1.1 and L1.2, with five sublineages, denoted L1.1.1, L1.1.2, L1.1.3, L1.2.1 and L1.2.2 and was subsequently expanded5. Recently, the scheme has been refined8 to consist of three main groups, L1.1-L1.3 and seven third-level sublineages, L1.1.1-L1.1.3 identical to the previous scheme, a novel L1.2.1, L1.2.2 equivalent to previous L1.2.1 and finally L1.3, equivalent to previous L1.2.2 and classified into L1.3.1 and L1.3.2.

Associations between each of these sublineages with geography have been studied. For example, L1.1.1.1 is predominant in Vietnam7, L1.1.3 is common in Myanmar8, Bangladesh9 and East Africa10 while L1.1.2.2 and L1.2.2 (previous L1.2.1) are associated with India and Southeast Asia, respectively5. Interestingly, around 80% of *Mtb* isolates in the Philippines, a high TB burden country, belonged to the last sublineage11. Thus, existing SNP-based classification schemes may not provide sufficient resolution for characterizing major sublineages in

1Faculty of Medicine, Siam University, Phet Kasem Road, Bangkok, Thailand. 2Pornchai Matangkasombut Center for Microbial Genomics, Department of Microbiology, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok, Thailand. 3Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. 4Department of Medical Sciences, Ministry of Public Health, Nonthaburi 11000, Thailand. 5Fukujuji Hospital and Research Institute of Tuberculosis, Japan Anti-Tuberculosis Association, Kiyose 204-8533, Japan. 6National TB Control Programme, Department of Public Health, Ministry of Health and Sports, Naypyitaw 15011, Myanmar. 7Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Had Yai 90110, Thailand. 8National Science and Technology Development Agency, Pathumthani, Thailand. 9email: Prasit.pal@mahidol.ac.th
results

phylogeny of L1 sublineages. WGS data of 1,764 M. tuberculosis (Mtb) L1 isolates from Asia, Oceania and Africa were obtained until November 2020. Most isolates were from Thailand (36%), Vietnam (26%), India (13%) and the Philippines (9%). The remaining were from other Asian countries (5%), Africa (n = 151; 9%) and Oceania (Australia and Hawaii; 2%) (Supplementary Table S1). Accession numbers of the raw paired-end reads of all Mtb isolates used in this study are provided in Supplementary Table S2.

The phylogeny inferred from whole-genome SNPs supported the previously-described three major clades, L1.1-L1.3, with subclades consistent with the existing schemes5,6 (Fig. 1).

L1.1 splits into three major sublineages, denoted L1.1.1–L1.1.3. Most (97%) of L1.1.1 isolates were from Thailand or Vietnam. L1.1.1.1 was the most common sublineage of L1.1.1 (51%) and accounted for 85% of the samples from Vietnam but only 3% of Thai samples. L1.1.1.4 was also mainly found in Vietnam while most of the remaining sublineages were predominantly found in Thailand. We designated two new clades in L1.1.1: L1.1.1.10 from Vietnam and Thailand and L1.1.1.11 from West Africa10. About 3% of L1.1.1 isolates were left as unclassified (Fig. 1 and Supplementary Table S3).

L1.1.2 splits into two deep-branching clades5. L1.1.2.1 consists entirely of Thai samples while L1.1.2.2 is widespread across Africa, South Asia, mainland Southeast Asia (MSEA), and is most common in India (80%) (Supplementary Table S3). Notably, L1.1.2.2 was not identified in samples from East Asia or island Southeast Asia (ISEA).

L1.1.3 contains four distinct clades, designated as L1.1.3.1-L1.1.3.4 (Fig. 1). The first three have been reported in many Asian studies5,8,13. L1.1.3.4 was identified mostly among African samples (Supplementary Table S3), particularly in Malawi (69%). Here, the fourth numbering of L1.1.3 clades is consistent with Palittapongarnpim et al.5 but differs from the Napiers’s scheme6, which does not recognize L1.1.3.3 (Fig. 1).

Interestingly, a sublineage in the Napiers’s scheme6 not present in our main dataset was L1.2.1, with most isolates being from patients in Europe. Therefore, we inferred a separate phylogeny of L1.2 isolates that included 410 isolates from our dataset and additional 364 isolates from Napier et al.6 and other recent studies10,14 (Fig. 2). L1.2 clearly splits into two major branches: L1.2.1 is a small clade (8.4% of L1.2) with mostly European samples while L1.2.2 is a large clade (91.6%) that is widespread across East Asia, MSEA and ISEA, consistent with Napier et al.6 (Table 2). Intriguingly, L1.2.1 contains a basal deep-branching clade of samples from East Timor and Papua New Guinea. L1.2.2 is classified into 5 sublineages, 1.2.2.1-1.2.2.5. L1.2.2.1-L1.2.2.3 were previously defined as L1.2.2.1-L1.2.2.3, respectively5. L1.2.2.1 was most common in Taiwan (31%), followed by the Philippines, Malaysia (Sabah) and Thailand (~10-15% each). L1.2.2.2 was mostly restricted to Thailand (89%). L1.2.2.3 was the largest sublineage (40% of L1.2). L1.2.2.4 was a small sister clade of L1.2.2.2. Both L1.2.2.3 and L1.2.2.4 were most common in the Philippines and nearby East Malaysia. L1.2.2.5 was the smallest sublineage, with most samples from Vietnam. The separation of L1.2.2 into five sublineages were supported by their comparable intragroup average pairwise SNP distances, which are less than the intergroup average pairwise SNP distances, and the fixation indices (Supplementary Fig. S2).

L1.3 (previously L1.2.2) was widespread in countries surrounding the Indian Ocean. It comprises L1.3.1 (18%) and L1.3.2 (82%) sublineages, with one basal isolate from Thailand. Almost all L1.3.1 isolates (88%) were from Eastern or Southern Africa, with a few basal isolates from India. In contrast, L1.3.2 was most common in India (36%), followed by Thailand (20%) (Supplementary Table S3).
Figure 1. Maximum-likelihood phylogeny of 1,764 L1 isolates. Highlighting and left column indicate sublineages used in this study, right column indicates Napier's scheme⁶. An interactive online version is available in Microreact at https://microreact.org/project/ii9NYi9YWaWFfZkiwV2joC.
Analysis of 68 DVRs and spoligotypes. All L1 isolates had characteristic deletions of six DVRs, DVR39–42, 44 and 48, in the CRISPR region. A small fraction (<3%) of L1.1.1 isolates in many sublineages had only those six deletions (Supplementary Table S4). The corresponding 77777777413371 (SIT236) spoligotype was expected to resemble that of the most recent common ancestor of L1 (L1-MRCA). As the CRISPR region evolves primarily by deletion, sublineages with more deleted DVR blocks are expected to be more derived. We found that L1.1.1 had the lowest number of deleted DVR blocks, 4 on average, followed by L1.3.2 (4.5), L1.1.2 (5), L1.1.3 (6), L1.2.2 (6) and L1.3.1 (7) (Supplementary Fig. S4).

Some sublineages correlated well to a unique set of DVR deletions that may be used as markers (Table 1). For example, L1.1.3 isolates share DVR33 and DVR56 deletions. L1.1.3.1 contained a clade with additional deletions of DVR6–7, DVR51 and DVR57, corresponding to spoligotype 777777547413771 (SIT292). L1.1.3.4 had additional deletions of DVR7–10, DVR12–19 and DVR34, corresponding to spoligotype 70077747413771 (SIT129). All L1.2 isolates shared DVR10 deletion. Most L1.2.1 had additional deletions of DVR11 and DVR33, except for a few basal isolates. Since DVR10–11 deletions were not used for spoligotyping, some L1.2.1 and L1.1.3 isolates do have the same spoligotype, 777777547413771 (SIT591), as a result of DVR33 deletion. Additional deletions of DVR4 and DVR30–31 were specific to L1.2.2 (typically 67000003413771, SIT89).

Several sublineages shared a single DVR deletion block, e.g. L1.1.1.5 (DVR18–21), L1.1.1.6 (DVR57–60) and L1.1.1.8 (DVR53) (Table 1 and Supplementary Fig. S1). Nevertheless, those deletions could occur in other sublineages as well, limiting their potential use as reliable markers. Deletions of some DVRs were clearly homoplastic, occurring several times independently in the evolutionary history of L1. DVR62 is a characteristic of both L1.3 and L1.1.9, and is deleted together with DVR61 in L1.1.1.10. This deletion was also found in many other sublineages of L1.1.1. Thus, DVR62 deletion by itself is of limited genotypic value.

Several DVR deletions appeared to occur sequentially along the phylogeny. For example, most of L1.1.1.1 (n = 349, 84%) belonged to a clade with shared DVR36–37 deletions. The rest (16%) had intact DVR36–37 and were more basal to this clade, with 16 isolates having only six DVR deletions similar to L1-MRCA (Supplementary Fig. S1 and Supplementary Table S4). Within the DVR36–37 deletion clade, there were two subclades with additional deletions of DVR23 (n = 25) or DVR29–63 (n = 35).

Extreme deletions resulted in the complete absence of DVRs in 7 isolates from Thailand, all belonging to L1.1.1.1 (n = 349, 84%) belonged to a clade with shared DVR36–37 deletions. The rest (16%) had intact DVR36–37 and were more basal to this clade, with 16 isolates having only six DVR deletions similar to L1-MRCA (Supplementary Fig. S1 and Supplementary Table S4). Within the DVR36–37 deletion clade, there were two subclades with additional deletions of DVR23 (n = 25) or DVR29–63 (n = 35).

Extreme deletions resulted in the complete absence of DVRs in 7 isolates from Thailand, all belonging to L1.1.1.1 (n = 349, 84%) belonged to a clade with shared DVR36–37 deletions. The rest (16%) had intact DVR36–37 and were more basal to this clade, with 16 isolates having only six DVR deletions similar to L1-MRCA (Supplementary Fig. S1 and Supplementary Table S4). Within the DVR36–37 deletion clade, there were two subclades with additional deletions of DVR23 (n = 25) or DVR29–63 (n = 35).

Extreme deletions resulted in the complete absence of DVRs in 7 isolates from Thailand, all belonging to L1.1.1.1 (n = 349, 84%) belonged to a clade with shared DVR36–37 deletions. The rest (16%) had intact DVR36–37 and were more basal to this clade, with 16 isolates having only six DVR deletions similar to L1-MRCA (Supplementary Fig. S1 and Supplementary Table S4). Within the DVR36–37 deletion clade, there were two subclades with additional deletions of DVR23 (n = 25) or DVR29–63 (n = 35).

Geographic diversity of L1. Sublineages of L1 distributed preferentially around the Indian Ocean and the Western Pacific region (Fig. 3). L1.2 was predominant in ISEA and was rarely reported from Africa while L1.1.1.11 was restricted to West Africa. Only L1.1.2.2 and L1.3 were distributed widely across Asia and Africa, particularly in South Asia.

At the country level, some countries, with >50 samples, had a distinct predominant L1 sublineage: L1.1.1.1 for Vietnam, L1.1.2.2 for India, L1.1.3.4 for Malawi, L1.2.2.1 for Taiwan, L1.2.2.3 for the Philippines and East Malaysia, L1.2.2.2 and L1.1.1.8 for Thailand (Fig. 3, Table 2, and Supplementary Table S3). India, Thailand and
Myanmar were the only countries in this dataset that harbored isolates belonging to all five common sublineages of L1, despite only 24 WGS samples from Myanmar.

**Sublineage barcoding SNPs.** We identified SNPs uniquely shared by all isolates within each sublineage as markers for sublineage identification (Supplementary Tables S5 and S6). Among the 1,835 sublineage-specific SNPs, ~80% were in non-essential coding regions or in non-coding regions. We recommend using the full set of SNP markers when possible. We also provided a subset of 125 barcoding SNPs, prioritizing synonymous SNPs within essential genes (Supplementary Table S6).

**Drug resistance mutations.** We used TBprofiler to predict drug resistance based on known genetic markers16 (Supplementary Fig. S5). The majority of isolates (n = 1,272, 72%) did not possess any resistance mutations.

### Table 1. Typical DVR deletions of some sublineages.

| Sublineage | n Typical DVR deletion | 10 | 20 | 30 | 40 | 50 | 68 | % with only the typical deletion | % of all isolates with typical deletion |
|------------|------------------------|----|----|----|----|----|----|-------------------------------|-------------------------------------|
| L1.1.1.1   | None                   |    |    |    |    |    |    | 4 N/A                         | 236 77777777413771                   |
|            | 36-37                  |    |    |    |    |    |    | 41 84                         | 139 7777774413771                   |
|            | 29-63                  |    |    |    |    |    |    | 7 8                           | 405 77777700000011                  |
| L1.1.1.2   | None                   |    |    |    |    |    |    | 3 N/A                         | 236 7777777413771                   |
|            | 5-6                    |    |    |    |    |    |    | 27 N/A                        | 139 7777774413771                   |
|            | 1-68                   |    |    |    |    |    |    | 21 N/A                        | 2669 00000000000000                 |
| L1.1.1.3   | None                   |    |    |    |    |    |    | 45 N/A                        | 236 7777777413771                   |
|            | 14-17                  |    |    |    |    |    |    | 18 N/A                        | 792 76377777413771                  |
| L1.1.1.4   | 51                     |    |    |    |    |    |    | 20 60                         | 234 77777777413771                  |
|            | 21                     |    |    |    |    |    |    | 12 16                         | 1372 77777777413771                 |
|            | 62                     |    |    |    |    |    |    | 12 16                         | 48 77777777413771                   |
|            | 5-9, 51                |    |    |    |    |    |    | 4 20                          | 234 77777777413771                  |
| L1.1.1.5   | 52 18-21               |    |    |    |    |    |    | 29 100                        | 951 774177777413771                 |
|            | 18-21, 59-62           |    |    |    |    |    |    | 25 37                         | 493 77417777741371                  |
| L1.1.1.6   | 38 57-60               |    |    |    |    |    |    | 24 100                        | 236 77777777413771                  |
| L1.1.1.7   | None                   |    |    |    |    |    |    | 21 N/A                        | 236 77777777413771                  |
|            | 34-54                  |    |    |    |    |    |    | 19 26                         | 1274 77777600000071                 |
| L1.1.1.8   | 100 53                 |    |    |    |    |    |    | 67 100                        | 256 77777777413671                  |
| L1.1.1.9   | 35 62                  |    |    |    |    |    |    | 29 100                        | 48 77777777413731                   |
| L1.1.1.10  | 10 61-62               |    |    |    |    |    |    | 90 100                        | 48 77777777413731                   |
| L1.1.1.11  | 10 4                   |    |    |    |    |    |    | 40 100                        | 342 67777777413771                  |
| L1.1.2     | 30 18                  |    |    |    |    |    |    | 33 100                        | 939 77577777413771                  |
|            | 18, 54-61              |    |    |    |    |    |    | 23 33                         | 939 77577777413771                  |
|            | 3-4, 50-53             |    |    |    |    |    |    | 39 71                         | 11 47777777413071                   |
| L1.1.3     | 37 6-7, 33, 51, 56-57  |    |    |    |    |    |    | 38 78                         | 292 7777775413371                  |
|            | 30 33, 56              |    |    |    |    |    |    | 63 100                        | 591 777777541371                   |
|            | 23 33, 56              |    |    |    |    |    |    | 70 100                        | 591 777777541371                   |
|            | 54 7-10, 12-19, 33-34, 56| | | | | | 87 100 | 129 70077777413771 |
| L1.2.1     | 65 10-11, 33           |    |    |    |    |    |    | 14 91                         | 591 777777541371                   |
| L1.2.2     | 152 4, 10, 30-31       |    |    |    |    |    |    | 68 100                        | 19 6777777413771                   |
|            | 147 4, 10, 17-35       |    |    |    |    |    |    | 78 100                        | 89 67400000341371                  |
|            | 311 4, 10, 30-31       |    |    |    |    |    |    | 69 100                        | 19 6777777413771                   |
|            | 84 4, 10, 30-31        |    |    |    |    |    |    | 32 100                        | 19 6777777413771                   |
|            | 13 4, 10, 30-31        |    |    |    |    |    |    | 54 100                        | 19 6777777413771                   |
| L1.3.1     | 26 13, 56-57, 60, 62   |    |    |    |    |    |    | 73 85                         | 48 77777777413731                  |
| L1.3.2     | 121 62                 |    |    |    |    |    |    | 54 98                         | 806 75777777413731                  |

Myanmar were the only countries in this dataset that harbored isolates belonging to all five common sublineages of L1, despite only 24 WGS samples from Myanmar.

**Sublineage barcoding SNPs.** We identified SNPs uniquely shared by all isolates within each sublineage as markers for sublineage identification (Supplementary Tables S5 and S6). Among the 1,835 sublineage-specific SNPs, ~80% were in non-essential coding regions or in non-coding regions. We recommend using the full set of SNP markers when possible. We also provided a subset of 125 barcoding SNPs, prioritizing synonymous SNPs within essential genes (Supplementary Table S6).

**Drug resistance mutations.** We used TBprofiler to predict drug resistance based on known genetic markers16 (Supplementary Fig. S5). The majority of isolates (n = 1,272, 72%) did not possess any resistance mutations.
Figure 3. Geographic distribution of 1,764 L1 sublineages across Asia, Africa and Oceania. Pie sizes are proportional to the total number of isolates from each location. The map was created using the R package ggplot2 v3.3.5 (https://ggplot2.tidyverse.org).

Table 2. Distribution of 774 L1.2 isolates by country and sublineage. The 410 L1.2 isolates from the main dataset were supplemented with 364 isolates from additional sources; see Supplementary Table S2.
Mutations associated with drug resistance were diverse. All rifampicin-resistant isolates had variants in the \( \text{rpoB} \) gene, with 31 distinct alleles associated with changes at 13 amino acid residues. For isoniazid resistance, 99% of the isolates (\( n = 344 \)) had mutations in \( \text{katG} \) or \( \text{fabG1/inhA} \) involving 36 alleles, with \( \text{katG} \) Ser315Thr being most common (52%), followed by the -15C > T mutation in the \( \text{embB} \) promoter (38%), with only 9 isolates having both variants (Supplementary Table S7). Among the other two first-line drugs, we identified 44 alleles of \( \text{embR} \) among 63 isolates conferring pyrazinamide resistance and 30 alleles in \( \text{embA} \), \( \text{embB} \) and \( \text{embR} \) conferring ethambutol resistance.

A small proportion of isolates (\( n = 28 \), 1.6%) possessed mutations in \( \text{gyrA} \) or \( \text{gyrB} \) conferring resistance to fluoroquinolones, with Ala90Val and Asp94(Gly/Ala/Asn) in \( \text{gyrA} \) being the most common (\( n = 26 \)). Among those isolates, 18 were qualified as pre-XDR and mostly belonged to L1.1.1 or L1.3.2. The only isolate identified as XDR belonged to L1.1.1.9 from Thailand. In addition to pre-XDR mutations, it also had an insertion in \( \text{Rv}0678 \) at position 192 that made it likely to resist bedaquiline.

**Genetic clustering.** We identified genetic clusters based on the number of SNP differences being within a pre-specified cut-point (Supplementary Fig. S7). Using three cut-points at 5, 12 and 20 SNPs, there were 20 (1.1%), 132 (7.6%) and 251 (14.4%) clustered isolates respectively, with an average cluster size of 2.0, 2.1 and 2.4 (Supplementary Table S8). The clusters were distributed across most sublineages, but were most common among L1.3 isolates (31%) and L1.1.3 (23%), and were rare in L1.1.1 (9.6%).

**Discussion**

Here, we provide an updated classification scheme of Mtib L1 based on phylogenetic analysis of WGS data of samples from L1 endemic regions. The major differences between our nomenclature and the Napier’s scheme are in L1.1.2, L1.1.3 and L1.2.2 sublineages (Fig. 1). Several sublineages associated with a specific country or region are recognized in our scheme but not in the Napier’s scheme. For instance, L1.1.2.1, L1.1.3.3, L1.2.2.2 and several sublineages of L1.1.1 are mainly associated with Thailand. We identified a small deep-branching clade of L1.2.1 with samples from New Guinea and East Timor, suggesting an underexplored endemic area of L1.2.1 in ISEA. It is possible that a more complex population structure remains to be revealed as more WGS data become available.

Comparisons between SNP sublineages and spoligotypes reveal some degree of congruence. There is greater congruence when the spoligotypes are defined by several deletion events, such as L1.1.3 and L1.2.2 (previously L1.2.1), which are equivalent to EA6_\text{BDG1} and EA12 respectively. In contrast, spoligotypes defined by a single deletion event are more likely to be homoplasmic. The most notable example is \( \text{SIT48} \), typically defined by the deletion of \( \text{DVR62} \), which occurred several times independently in the evolutionary history of L1. \( \text{SIT48} \) can be a member of L1.1.1 or L1.3. Fortunately, since L1.1.1 is largely restricted to MSEA, \( \text{SIT48} \) in other regions, such as India or Africa, most likely corresponds to L1.3 whereas \( \text{SIT48} \) in MSEA may belong to either L1.1.1 or L1.3 as shown in Fig. 4.

We also clarify several genetic relationships between spoligotypes that could facilitate the development of better genetic groupings of spoligotypes. This is important as spoligotyping is more affordable in high-burden countries and provides data from much larger sample sizes with a higher geographic coverage than WGS.

The geography of isolates with most intact DVRs, similar to L1-MRCA, could be indicative of the region where L1 originated. Such isolates have the \( \text{SIT236} \) spoligotype and are found exclusively in certain sublineages of L1.1.1, except for two L1.3.2 isolates (Fig. 4 and Supplementary Table S4). Since all sublineages of L1.1.1 except for L1.1.1.11 are mostly restricted to MSEA, where the diversity of L1 is also greatest (Fig. 3), it is possible that the L1-MRCA might have been present there. This is in contrast with recent genomic studies which predicted the origin of L1 in South Asia, where none of the \( \text{SIT236} \) isolates were present in our dataset. We note that their inferences were dominated by L1.1.2.2 and L1.3, and may suffer from small sample sizes (~300 or lower) and non-representative sampling.

One major limitation of this study is that it was based on WGS data collected for various purposes. Epidemiological findings such as the prevalence of drug resistance and genetic clusters were therefore only suggestive and require further investigation. A few source studies were intended for drug resistant bacteria which biased the estimates of drug resistance frequencies.

Our analysis revealed a vast diversity of resistance-conferring mutation combinations. The results confirmed the presence of the mutations in the rifampin-resistance determining regions (RRDR) in the vast majority (95%) of rifampin resistance isolates, which would allow the GeneXpert system to adequately detect rifampin resistance. Although available commercial line probe assays do not specifically identify mutations in about 37% of...
rifampicin-, 13% of isoniazid- and 55% of ethambutol-resistant isolates in this dataset, they should be able to
diagnose the resistance by the absence of hybridization of wild-type probes. The line probe assay specifically
detects most fluoroquinolone resistance mutations.20

Conclusion
This study refines a genotyping scheme of L1, particularly of the L1.2 sublineages. Many basal isolates of L1.2
were recognized in ISEA samples, suggesting a possible study site to further elucidate the evolutionary history
of the clade. Mapping spoligotypes onto the SNP phylogeny clarifies the genotypic identity of some homoplastic
spoligotypes. Analysis of drug resistance mutations revealed a vast diversity of mutated alleles, particularly
among the first-line drugs. This has implications for the utility of genetic tests for diagnosis of drug-resistant
TB. Finally, genetic clusters in L1 are rare and small in size, suggesting a more limited transmission potential
of L1, especially in comparison with L2.21

Methods
Dataset. We compiled a collection of whole-genome sequencing (WGS) data of 1,764 M. tuberculosis
isolates from countries in Asia, Oceania and Africa. This is the main dataset. Most of the data were publicly
available while a small proportion was from unpublished studies. Supplementary Table S1 provides a list of data
sources and references for the WGS data used in this study. A full list of accession numbers for all 1,764 isolates
is in Supplementary Table S2. Supplementary Table S3 shows the distribution of samples across countries and
sublineages. We also obtained additional 364 L1.2 isolates from three recent studies6,10,14. This set was analyzed
together with the 410 L1.2 isolates from the main dataset, referred to as the L1.2 dataset (n = 364 + 410 = 774)
(Supplementary Table S2).

SNP calling. Short reads data were quality-trimmed using trimmomatic v0.39 (sliding-window trimming
with window size of 4 and read quality threshold of 30)22. The processed short reads were mapped to the
M. tuberculosis H37Rv reference genome (NC_000962.3) using bwa mem23. Duplicate reads were masked using
Picard’s MarkDuplicates, and per-sample SNPs were called using GATK HaplotypeCaller v4.1.6.0 (base quality
score ≥ 20, haploid model)24.

As additional sample quality control, we excluded samples if (i) the median read depth below 10 or the median
breadth of coverage (depth at least 10) below 10%, or (ii) it was redundant as determined by the BioSample acces-
sion number and sample metadata if provided by the original study, or (iii) the processed reads had unusual
%GC content, or (iv) SNPs specific to different sublineages were present in substantial proportions, indicative
of mixed strains. The total of 1,764 isolates passed this sample quality control.

Joint SNP calling of all 1,764 isolates was performed using GATK GenotypeGVCFs v4.1.6.024. For variant
quality control, we dropped SNPs within a region annotated as repeat region, IS6110, PE, PPE, mobile element,
phage or 13E12 repeat family. SNPs within known drug-resistant genes were excluded as in Ajawatanawong
et al25. Finally, SNPs with quality by depth (QD) < 2 or root mean square of mapping quality (MQ) < 40 were
also excluded. This resulted in 153,244 SNPs across 1,764 genomes. The additional L1.2 isolates were processed
similarly.

Phylogenetic analysis. SNP calls passing the quality filters were converted into multiple sequence align-
ment using a custom script based on GATK VariantsToTable (available at https://github.com/CENMIG/snppl
et). Point deletions and missing calls were converted into gaps in the alignment. The H37Rv strain (L4) was
used as an outgroup. A maximum likelihood (ML) phylogeny was inferred using IQ-TREE v2.1.128. The best-
fit substitution model was determined to be GTR + G4 by ModelFinder27. The bootstrap branch supports were
based on 1,000 replicates. We also inferred an ML phylogeny correcting for ascertainment bias as shown in Sup-
plementary Fig. S8 and obtained consistent bootstrap support for all sublineages29.
Population structure. Summary statistics of genetic diversity of each sublineage and genetic divergence between sublineages were calculated from the filtered SNPs using scikit-allele v1.3.129. These included genome-wide averaged nucleotide diversity (π) for each sublineage, pairwise SNP difference and pairwise FST (using Hudson estimator30) for within sublineage and between sublineages, and principal component analysis (PCA) of genome-wide SNPs.

Detecting direct variable repeat (DVR) deletions. To determine deletion pattern of the 68 known DVRs, we mapped the short-read data to a custom-made reference sequence that contained all the DVRs. This reference was constructed based on an L1 isolate (accession: ERR752247) with 62 DVRs, augmented with DVRs 39–42 from the H37Rv strain (accession: NC_000962.3) and DVRs 44 and 48 from an L2.1 isolate (accession: ERR718276).

Each DVR segment was identified as a deletion if one of the following criteria was satisfied: (i) the median read depth was below a cutoff and the proportion of sites with at least one read mapped to was < 50%; (ii) the 25% quantile of read depths was < 3; (iii) the 5% quantile of read depths was 0; (iv) if the overall median depth (across all DVRs) was > 50, the median read depth was smaller than a half of the median read depth. For (i), we used the cutoff value of 5, or half of the baseline median read depth calculated from all DVR regions, whichever was lower. This choice as well as the criteria (ii) and (iii) accounted for low-depth samples. The criterion (iv) accounted for samples with high sequencing depths by adjusting for overall read depth across all DVR regions. Given a DVR pattern, we extracted a relevant set of 43 DVRs to determine the spoligotype. SIT and spoligotype group for each spoligotype was obtained from the SITVIT2 database31. The code for DVR calling is available in GitHub at https://github.com/ythawn/dvr caller.

We tested our algorithm using a 480-isolate subset where spoligotypes had been determined experimentally and DVR patterns had been independently identified using a de novo assembly-based approach7. The above criteria yielded ~ 6% difference for the spoligotype and 16% difference for the DVR pattern. We also compared our predicted spoligotypes with two other methods, SpoTyping32 and Galru33 (Supplementary Table S2). The number of mismatches between each pair of methods ranged between 14%–17%. The difference in predicted spoligotypes was usually only about 1–2 DVR segments. Manual inspection of mapped read depths indicated that our method appeared to be more reliable.

Revising genotyping scheme and identifying lineage-specific SNPs. We used the following criteria to delineate genotypes based on the inferred phylogeny (Fig. 1). First, we preserved the genotypes that have previously been defined. Sublineage of each isolate was identified using two existing SNP-based genotyping schemes for M. tuberculosis L15,5. Second, we assigned a new genotype to a monophyletic clade of previously unclassified strains with at least 10 isolates, bootstrap support ≥ 90%. We also required that all isolates shared at least one SNP specific to the group that was absent outside the group. Third, we expanded the definition of existence of unclassified strains with at least 10 isolates, bootstrap support ≥ 90%. We also required that all isolates shared at least one of bedaquiline or linezolid35,36.

Summary statistics of genetic diversity of each sublineage and genetic divergence between sublineages were calculated from the filtered SNPs using scikit-allele v1.3.129. These included genome-wide averaged nucleotide diversity (π) for each sublineage, pairwise SNP difference and pairwise FST (using Hudson estimator30) for within sublineage and between sublineages, and principal component analysis (PCA) of genome-wide SNPs.

Drug resistance mutations. We used TBProfiler16 to predict resistance of each isolate to the following 19 anti-TB drugs: rifampicin, isoniazid, pyrazinamide, ethambutol, streptomycin, moxifloxacin, ofloxacin, levofloxacin, ciprofloxacin, amikacin, kanamycin, capreomycin, ethionamide, para-aminosalicylic acid, cycloserine, linezolid, bedaquiline, clofazimine and delamanid. Classification of drug resistance type was according to the latest World Health Organization (WHO) recommendations as follows. mono-resistance (mono-DR): one first-line drug (isoniazid, rifampicin, ethambutol, pyrazinamide); polypredrug resistance (poly-DR): more than one first-line drug other than isoniazid and rifampicin together; multidrug resistance (MDR): both isoniazid and rifampicin together; pre-XDR: MDR and any fluoroquinolone; extensive drug resistance (XDR): pre-XDR and at least one of bedaquiline or linezolid35,36.

Genetic clustering. We identified genetic clusters as clades in the phylogeny containing isolates that can be linked via pairwise SNP distances of at most 5, 12 or 20 SNPs37,38.

Data availability Accession numbers of the raw paired-end reads of all Mtb isolates used in this study are provided in Supplementary Table S2.

Received: 24 August 2021; Accepted: 11 January 2022
Published online: 28 January 2022

References
1. Comas, I. et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet 45, 1176–1182. https://doi.org/10.1038/ng.2744 (2013).
Author contributions
T.N., Y.T., P.P. designed study. S.M., H.Y., H.M.W.M. and V.C. provided sequence data. T.N. performed analysis, and drafted manuscript. Y.T. collected W.G.S. data, performed analysis, and drafted manuscript. PP conceptualized, drafted and finalized manuscript. All authors reviewed the paper.
Funding
This work was supported by Mahidol University Integrated and Multidisciplinary Research Cluster grant, the Emerging Infectious Disease program, National Science and Technology Development Agency, Department of Medical Sciences, Ministry of Public Health and the Fogarty International Center, the Fogarty International Center and the National Institute of Allergy and Infectious Diseases, of the National Institutes of Health under Award Number D43 TW009522. The content is solely the responsibility of the authors and does not necessarily represent the official views of funders.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-05524-0.

Correspondence and requests for materials should be addressed to P.P.

Reprints and permissions information is available at www.nature.com/reprints.

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

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

© The Author(s) 2022