

# SCIENTIFIC REPORTS

OPEN

## Evidence for Host-Bacterial Co-evolution via Genome Sequence Analysis of 480 Thai *Mycobacterium tuberculosis* Lineage 1 Isolates

Prasit Palittapongarnpim<sup>1,2</sup>, Pravech Ajawatanawong<sup>1</sup>, Wasna Viratyosin<sup>2</sup>, Nat Smittipat<sup>2</sup>, Areeya Disratthakit<sup>3</sup>, Surakameth Mahasirimongkol<sup>3</sup>, Hideki Yanai<sup>4,5</sup>, Norio Yamada<sup>6</sup>, Supalert Nedsuwan<sup>7</sup>, Worarat Imasanguan<sup>7</sup>, Pacharee Kantipong<sup>7</sup>, Boonchai Chaiyasirinroje<sup>4</sup>, Jiraporn Wongyai<sup>4</sup>, Licht Toyo-oka<sup>8</sup>, Jody Phelan<sup>9</sup>, Julian Parkhill<sup>10</sup>, Taane G. Clark<sup>9</sup>, Martin L. Hibberd<sup>9</sup>, Wuthiwat Ruengchai<sup>1</sup>, Panawun Palittapongarnpim<sup>1</sup>, Tada Juthayothin<sup>2</sup>, Sissades Tongshima<sup>2</sup> & Katsushi Tokunaga<sup>8</sup>

Tuberculosis presents a global health challenge. *Mycobacterium tuberculosis* is divided into several lineages, each with a different geographical distribution. *M. tuberculosis* lineage 1 (L1) is common in the high-burden areas in East Africa and Southeast Asia. Although the founder effect contributes significantly to the phylogeographic profile, co-evolution between the host and *M. tuberculosis* may also play a role. Here, we reported the genomic analysis of 480 L1 isolates from patients in northern Thailand. The studied bacterial population was genetically diverse, allowing the identification of a total of 18 sublineages distributed into three major clades. The majority of isolates belonged to L1.1 followed by L1.2.1 and L1.2.2. Comparison of the single nucleotide variant (SNV) phylogenetic tree and the clades defined by spoligotyping revealed some monophyletic clades representing EAI2\_MNL, EAI2\_NTM and EAI6\_BGD1 spoligotypes. Our work demonstrates that ambiguity in spoligotype assignment could be partially resolved if the entire DR region is investigated. Using the information to map L1 diversity across Southeast Asia highlighted differences in the dominant strain-types in each individual country, despite extensive interactions between populations over time. This finding supported the hypothesis that there is co-evolution between the bacteria and the host, and have implications for tuberculosis disease control.

The burden of tuberculosis is high in developing countries in Asia, with 56% of all global cases found in five countries: India, Indonesia, the Philippines, Pakistan and China<sup>1</sup>. With tremendous efforts, some new drugs have recently been developed, but there has still been limited success in developing new effective tuberculosis vaccines. Thus a better understanding of the interactions between *M. tuberculosis* and its host is urgently needed.

*M. tuberculosis* lineages strongly associate with the geographic location, ethnicity and ages of the hosts<sup>2</sup>, which could be due to the founder effect or co-adaptation between the bacterial and host population, favouring their stable coexistence<sup>3,4</sup>. The latter hypothesis was supported by the sympatric association between bacterial lineages and patients' origins<sup>5</sup> which was weakened in HIV-infected patients<sup>6</sup>. The global population structure of *M. tuberculosis* lineage 4 further supported the co-evolution hypothesis<sup>7</sup>. Moreover, some candidate gene studies revealed

<sup>1</sup>Department of Microbiology, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok, Thailand. <sup>2</sup>National Centre for Genetic Engineering and Biotechnology, National Science and Technology Development Agency, Phahonyothin Road, Pathumthani, Thailand. <sup>3</sup>Department of Medical Sciences, Ministry of Public Health, Tiwanon Road, Nonthaburi, Thailand. <sup>4</sup>TB-HIV Research Foundation, Chiangrai, Thailand. <sup>5</sup>Fukujuji Hospital, Japan Anti-tuberculosis Association (JATA), Kiyose, Japan. <sup>6</sup>Research Institute of Tuberculosis, JATA, Kiyose, Japan. <sup>7</sup>Chiangrai Prachanukroh Hospital, Ministry of Public Health, Chiangrai, Thailand. <sup>8</sup>Department of Human Genetics, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan. <sup>9</sup>London School of Hygiene and Tropical Medicine, London, UK. <sup>10</sup>Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. Correspondence and requests for materials should be addressed to Pr.P. (email: [Prasit.pal@mahidol.ac.th](mailto:Prasit.pal@mahidol.ac.th))

an association with diseases caused by only some lineages<sup>8</sup>. Recently, genome-wide association studies (GWAS) successfully identified the HLA-DRB1\*09:01 and CD53 genes associated with tuberculosis when the bacterial genotypes were considered<sup>9,10</sup>.

Other aspects of genotype-phenotype associations have also been studied, particularly for lineages 2 and 4. The association varied between sublineages. The sublineages of the Beijing strains were associated with differential survival rates, virulence and responses to vaccine<sup>2,11–13</sup>.

Unfortunately, there have been few studies on the phenotypes of lineage 1 (L1). L1, as defined by SNP typing, is congruent with the Indo-oceanic strains, defined by LSP (large sequence polymorphisms)<sup>14</sup> and almost completely congruent with the EAI (East African Indian) strains, defined by spoligotyping. The EAI strains were commonly reported in countries around the Indian Ocean<sup>15,16</sup> and associated with higher ages of patients<sup>17</sup>. Pulmonary tuberculosis caused by L1 strains was more likely to be sputum negative and had a higher sputum conversion rate at two months after treatment<sup>18</sup>. L1 was less associated with drug resistance than L2<sup>19,20</sup>. They were occasionally reported to associate with extrapulmonary tuberculosis<sup>21</sup>. In general, L1 isolates induced stronger cytokine responses and grew slower in macrophage culture<sup>22,23</sup>.

Studies of L1 are essential for the End TB strategy<sup>24</sup>. It is endemic in 11 of 30 high-burden countries. Some bacterial isolates, such as the “Asian human” type, from India and East Africa had distinct properties<sup>25</sup> and were hypothesized to provide an explanation for the lack of efficacy of the BCG vaccine in the monumental clinical trial in Tamil Nadu, Southern India<sup>26</sup>. The described phenotypes might associate with L1 strains.

Recently, Coll *et al.* proposed classification based on a whole-genome single nucleotide variant (SNV) phylogenetic tree and provided a corresponding barcoding scheme<sup>27</sup>. While there have been many WGS studies of lineage 2 strains, studies of L1 strains are rather limited.

Comparing epidemiological findings between countries requires a reliable and high-resolution classification based on the bacterial genetic relationship. EAI isolates are classified by spoligotyping into “clades”, such as EAI1–EAI8. However, it is not clear whether the “clades” are monophyletic. There are “orphan” spoligotypes that are not yet assigned to any clades while assignment of some spoligotypes may be problematic. Correlation of the SNV phylogenetic tree to spoligotypes would provide better insights into the phylogeography of L1.

In this study, the genomes of 480 isolates of *M. tuberculosis* L1 in Chiangrai, the northernmost province of Thailand, were sequenced and revealed significant diversity. Analysis of WGS was done to refine the SNV phylogenetics of L1, verify spoligotype clades and provide mapping between spoligotypes and the SNV phylogenetic tree. This finding provided a basis to map the distribution of predominant sublineages of L1 in various countries in Southeast Asia. Our results suggested an association between predominant sublineages and geography and ethnicity, despite a history of extensive interactions between various ethnic groups, supporting the hypothesis that there is some level of co-adaptation between *M. tuberculosis* L1 and human hosts.

## Results

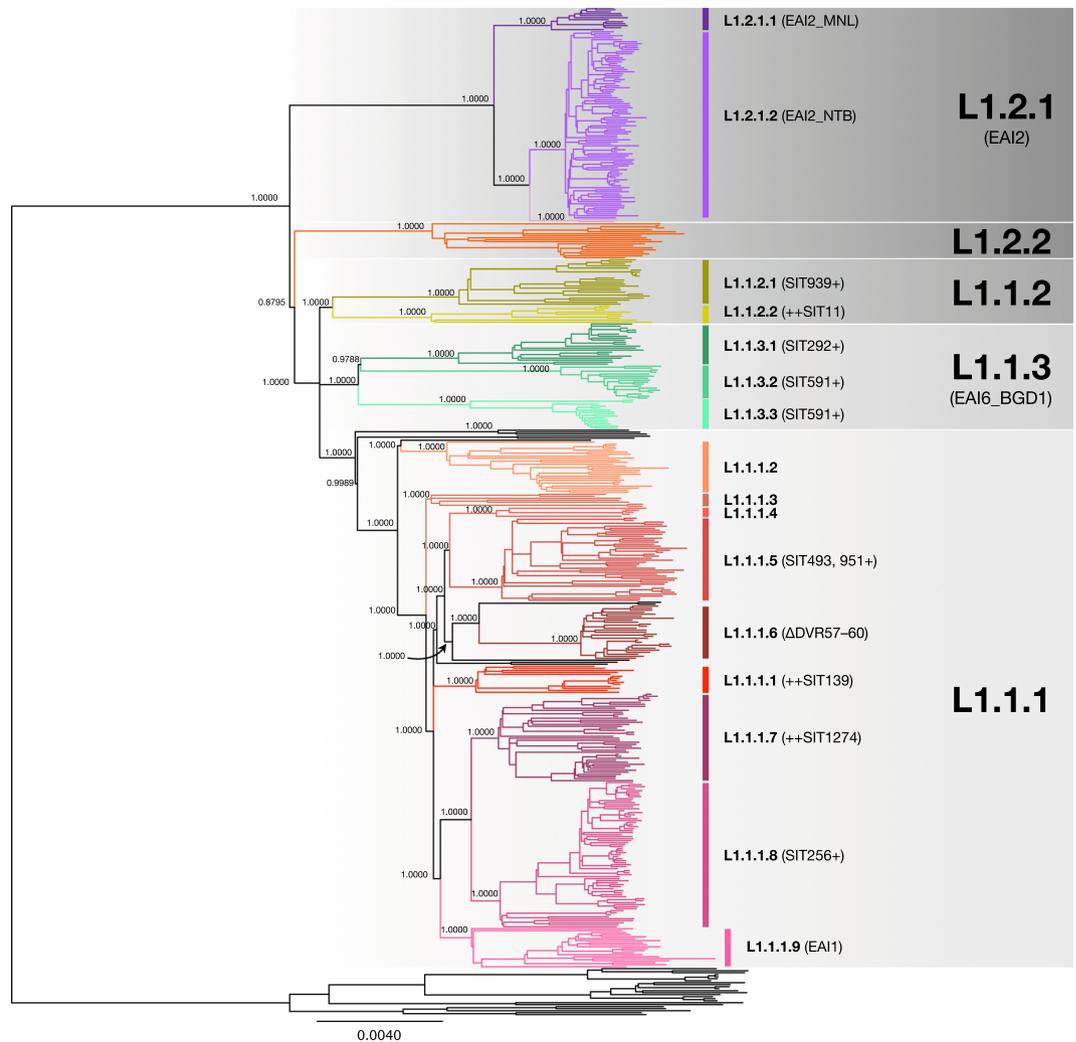
**Classification of *M. tuberculosis* L1.** Among 1174 successfully sequenced *M. tuberculosis* isolates, 480 were initially identified by LSP as belonging to the Indo-Oceanic family and had the 330 SNPs, previously reported to be specific to L1<sup>27</sup> (Supplementary Table S1). The following results describe phylogenetic analysis of WGS information from the 480 isolates. The resulting sublineages were then compared to the sequences of the Direct Repeat (DR) region obtained from WGS and examined by spoligotyping. The DR region contains a maximum of 69 DR sequences, interposed by 68 Direct Variable Repeats (DVRs). Variations between strains are due to the absence of some DVRs. The information on the absence of each of the 68 DVRs was extracted from WGS. Standard spoligotyping determines the absence of 43 selected DVRs and usually reports results for each isolate as a 15-digit octal code or equivalently designated spoligotype international type (SIT).

A total of 41,157 SNVs were identified, with 22,491 SNVs, each present in only a single isolate. The total identified SNVs were more than in the previous reports (Supplementary Table S2). L1 is the most genetically diverse lineage of *M. tuberculosis*<sup>2,27</sup>. Its mean pairwise SNV distance (MPSD) in this study was 645.6 (range 0–965), while that of the globally reported L1 isolates was 730<sup>2</sup>, indicating high heterogeneity of L1 in Chiangrai. This finding might be related to the knowledge that Chiangrai is an ancient settlement originally with Austroasiatic language-speaking people and subsequently replaced by people who speak the Tai-Kadai language family<sup>28,29</sup>. It was historically controlled by several tribal kingdoms that resided now in Thailand, Myanmar and Lao. Chiangrai is home to several hill tribes and recent Chinese immigrants. Its modern roles as a transport hub and a major tourist attraction site lead to rapid urbanization and migration from throughout Thailand and other countries.

The spoligotypes of 410 (85%) isolates were classified as EAI strains, while those of the other 66 isolates were not previously described<sup>30</sup>. However, one and three isolates had spoligotypes of 77777777773771 (SIT100) and 7777777777771 (SIT523) respectively. Neither pattern belongs to the EAI spoligotypes even though all four had the genetic markers of L1 (TbD1+ and RD239–). Isolates with SIT100 belonging to L1 have been previously reported<sup>31</sup>.

Analysis of all 68 DVR segments revealed that all EAI isolates and unclassified isolates lacked DVR39–42, DVR44, known characteristics of EAI, and DVR48, which is not used for standard spoligotyping. This finding indicated that all unclassified isolates belonged to the EAI family. Some isolates had all the other 62 DVRs, indicating that the most recent common ancestor of the EAI group (EAI-MRCA) carried the deletion of only DVR39–42, 44 and 48.

**Whole-genome SNV phylogenetic tree of L1.** Coll *et al.* studied 1601 globally collected WGS deposited in public databases, classifying 121 L1 isolates into five major groups, 1.1.1–1.1.3 and 1.2.1–1.2.2<sup>27</sup>. The phylogenetic trees of the 480 isolates constructed by Bayesian inference (Fig. 1) and maximum likelihood (Supplementary Fig. S1) methods were very similar to each other and mostly conformed to Coll’s report. They revealed deep branching into three major groups, corresponding to L1.1, L1.2.1 and L1.2.2, as indicated by the similarity of

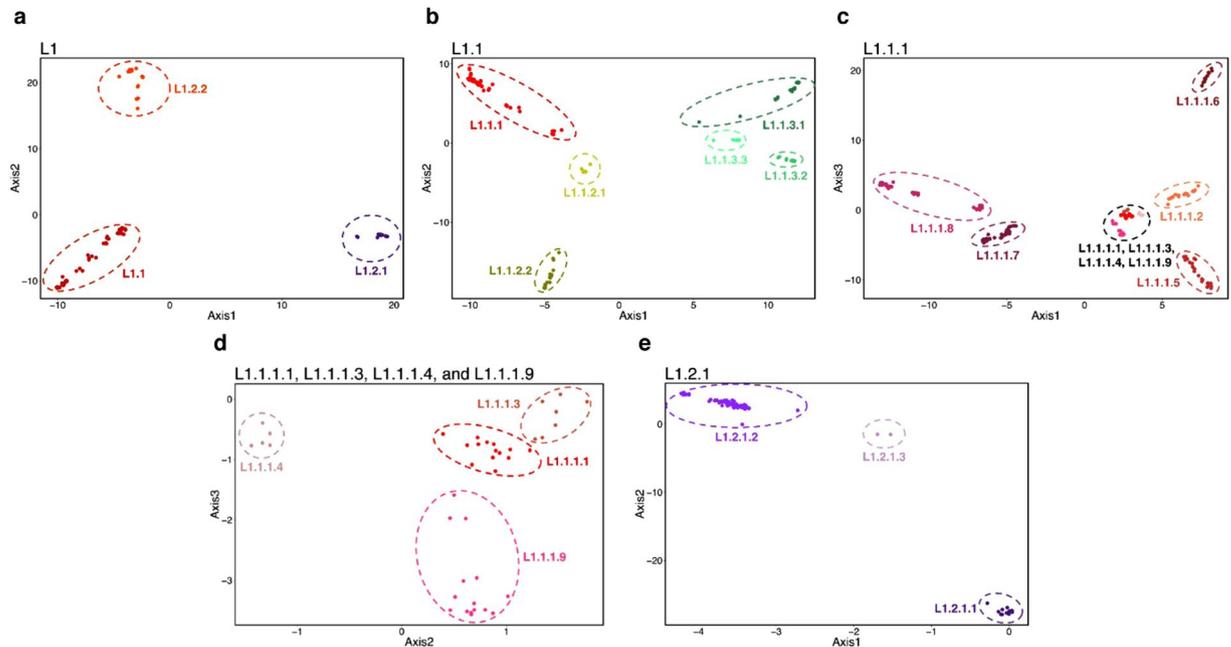


**Figure 1.** The phylogenetic tree of 480 *M. tuberculosis* L1 isolates from Chiangrai, constructed by the Bayesian inference method. The major subgroups were labelled on the rightmost. L1.1.1.1 were the same group as reported by Coll *et al.* The sublineages L1.1.1.2–L1.1.1.9, labelled on the right, were named sequentially according to their positions in the tree. Some interesting spoligotypes are shown after the sublineage names. The SIT number followed by the + sign indicates that the shown SIT contributed to the majority of the sublineage. The ++ sign before the SIT number indicates that the spoligotype was only a small fraction of the sublineage. The deletion of DVR57-60 was specific to L1.1.1.6 but was not revealed by spoligotyping because the DVRs were not used for standard spoligotyping.

many sublineage-specific SNPs in Coll's work and in our study (Supplementary Table S1). The recommended barcoding SNP for L1.2.1 (3479545 C/A), however, was not found.

The topological relationships between the three groups were different. Here, L1.2.2 appeared slightly more related to L1.1, but not L1.2.1, conforming to the presence of isolates with a single copy of *IS6110* only in L1.1 and L1.2.2 but not in L1.2.1 (Supplementary Table S3). The MPDs between 1.1–1.2.1, 1.1–1.2.2 and 1.2.1–1.2.2 were 800.8, 853.6 and 846.1 respectively, which contributed to the major peak (peak A) in the frequency distribution curve of the pairwise SNV distances, as shown in Supplementary Fig. S2. The nearly equal distances indicated that L1 could be divided into three distinct sublineages, as also shown by the PCA plot in Fig. 2a. The numbers of isolates belonging to the three sublineages were 354 (73.75%) for L1.1, 108 (22.5%) for L1.2.1 and 18 (4.75%) for L1.2.2. Because the number of L1.2.2 isolates was less than in Coll's report<sup>27</sup> and all were from only Thailand, the topology of our trees must be confirmed and the name L1.2.2 was maintained throughout this paper.

L1.1 itself comprises three further sublineages, 1.1.1–1.1.3, with all the sublineage-specific SNPs herein previously identified in Coll's report<sup>27</sup>. The separation was supported by the bootstrap score of 100%, PCA (Fig. 2b) and the fixation indices between 0.33–0.4 (Table 1). The MPDs between the three groups were approximately 750 (Supplementary Table S4 and Supplementary Fig. S2). The proportions of isolates belonging to these three sublineages were 76.0% (269) for L1.1.1, 9.0% (32) for L1.1.2 and 15.0% (53) for L1.1.3, which differed considerably from Coll's report. The recommended specific barcoding SNPs for both L1.1.2 and L1.1.3 were not identified herein, suggesting the presence of previously unidentified subgroups in this study. The recommended barcoding SNP for L1.1.2<sup>27</sup> was specific to a subset, L1.1.2.2, instead, indicating that L1.1.2.1 isolates were not included in Coll's study.



**Figure 2.** PCA plots revealed the separation of sublineages. (a) shows a plot of representative isolates of L1 with axes representing eigenvector 1 and 2. The red, purple and orange dots indicate isolates belonging to L1.1, L1.2.1 and L1.2.2 respectively. (b) shows a plot of representative isolates of L1.1 with axes representing eigenvector 1 and 2. Members of L1.1.1 are represented in red. The sublineages of L1.1.2 and L1.1.3 are indicated in different shades of greyish green and green, respectively. (c) shows a plot of representative isolates of L1.1.1 with axes representing eigenvector 1 and 3. (d) shows a plot of representative isolates of L1.1.1.1, L1.1.1.3, L1.1.1.4 and L1.1.1.9 with axes representing eigenvector 2 and 3. (e) shows a plot of representative isolates of L1.2.1 with axes representing eigenvector 1 and 2. L1.2.1.1, L1.2.1.2 and L1.2.1.3 are indicated in dark purple, blue and light purple, respectively.

The presence of the other peaks corresponding to smaller distances in the frequency distribution curve of pairwise SNV distances (Supplementary Fig. S2), as well as the large number of isolates, suggested the possibility for further sub-classification. In total, 18 sublineages were tentatively identified, with nine of them in 1.1.1, two in 1.1.2, three each in 1.1.3 and L1.2.1, and one in 1.2.2. The information regarding each sublineage is summarized in Tables 1 and 2. The classification is supported by the phylogenetic trees, with each sublineage having a bootstrap score of 100% and a posterior Bayesian probability of 1.0, and the results of the PCA plots (Fig. 2c–e). This classification is also supported by the statistically significant differences between the  $\pi_{\text{within group}}$  and  $\pi_{\text{between group}}$  (Wilcoxon rank-sum test with Bonferroni correction,  $\alpha < 0.00278$ ) or the fixation indices greater than 0.33 (Table 1). Only L1.1.1.4 and L1.2.1.3 had high fixation indices but insignificant differences between  $\pi_{\text{within group}}$  and  $\pi_{\text{between group}}$ , probably due to the low sample numbers. L1.1.1.3 had a low fixation index but statistically significant difference between  $\pi_{\text{within group}}$  and  $\pi_{\text{between group}}$ . The inter-group MPSDs of sublineages of L1.1.1 were mostly around 450–550 (Supplementary Table S4) and contributed to peak B in Supplementary Fig. S2. Eleven isolates were not classified into any groups. Numbering of the sublineages of L1.1.1 was performed so that L1.1.1.1 was the same as in Coll's assignment<sup>27</sup>, while the others were numbered sequentially according to their positions in the Bayesian tree. The MPSDs between each sublineage are shown in Supplementary Table S4.

This classification should provide a framework for identifying more variants belonging to L1 as well as for comparison of WGS across countries. The SNPs specific to each sublineage are provided in Supplementary Table S1. It is conceivable that the specific SNPs for each sublineage may decrease when more samples are examined.

As indicated by the intragroup MPSD, the diversity within each sublineage varied, which may be related to the level of transmission in Chiangrai. A smaller MPSD indicates a closer relationship between members of the sublineage, which may result from higher transmission activity, either in Chiangrai or in the vicinity. A large MPSD indicates that most isolates in the sublineage are distantly related and might have been introduced into Chiangrai separately. For L1.1.3, both events might occur. The MPSD of L1.1.3 was approximately three times higher than that of its three sublineages, suggesting that there were at least three clades belonging to L1.1.3, each of which was probably separately introduced to Chiangrai.

**Concordance between the SNV phylogenetic tree and clades defined by spoligotypes.** The identification of DR regions from the DNA contigs was successful in 476 of 480 isolates. Unfortunately, the identification of DR regions in the non-EAI isolates, belonging to SIT100 and SIT513, was unsuccessful. A comparison of the identified DVRs with experimental spoligotypes revealed 23 cases (4.8%) of discrepancy. In seven cases,

Sublineages	Number of isolates	Number of specific SNVs	Shared DVR deletions	Coll's classification	Number of specific SNPs in Coll's study	Common specific SNPs in both studies	Confirmed barcoding SNPs	Mean distances within sublineages /between sublineages	Probability	Fixation Indices	Depths/ heights in the phylogenetic tree
<b>I</b>	480	550		<b>1</b>	473		615938 G/A				0/23
<b>I.1</b>	354	23		<b>1.1</b>	38	22	4404247 G/A	570.0/808.4.1	$<2.2 \times 10^{-16}$	0.295	2/21
<b>I.1.1</b>	269	39		<b>1.1.1</b>	57	39	3021283 G/A	451.1/749.7	$<2.2 \times 10^{-16}$	0.398	3/20
<b>I.1.1.1</b>	14	44		<b>1.1.1.1</b>	138	44	3216553 G/A	252.9/468.3	$7.45 \times 10^{-6}$	0.460	8/8
I.1.1.2	26	42						340.5/554.2	$6.53 \times 10^{-10}$	0.386	7/10
I.1.1.3	7	5						431.8/503.7	0.002331	0.143*	7/3
I.1.1.4	5	47						224.2/477.2	0.01167*	0.530	11/4
I.1.1.5	42	46	18–21					323.9/505.5	$3.12 \times 10^{-15}$	0.359	11/11
I.1.1.6	27	90	57–60					135.3/490.6	$3.02 \times 10^{-10}$	0.724	10/11
I.1.1.7	44	29						269.8/453.4	$6.79 \times 10^{-16}$	0.405	10/12
I.1.1.8	73	34	53					174.4/467.0	$2.2 \times 10^{-16}$	0.627	10/13
I.1.1.9	20	23	62 (not unique)					284.7/487.3	$1.45 \times 10^{-11}$	0.416	9/12
Unclassified	11										
<b>I.1.2</b>	32	12		<b>1.1.2</b>	154	12		487.5/750.7	$6.49 \times 10^{-12}$	0.351	3/9
I.1.2.1	23	143	18					322.0/720.5	$6.60 \times 10^{-99}$	0.553	4/8
I.1.2.2	9	102	3–4				2622402 G/A**	310.0/720.5	$4.11 \times 10^{-5}$	0.570	4/5
<b>I.1.3 (EAI6)</b>	53	29	33, 56	<b>1.1.3</b>	66	29	—	502.1/750.2	$<2.2 \times 10^{-16}$	0.331	3/11
I.1.3.1	21	85	6–7, 33, 51, 56–57					228.6/654.6	$3.12 \times 10^{-8}$	0.651	5/10
I.1.3.2	17	215	33, 56					131.4/671.9	$6.99 \times 10^{-7}$	0.805	5/9
I.1.3.3	15	125	33, 56					138.4/652.3	$3.38 \times 10^{-6}$	0.788	4/9
<b>I.2.1 (EAI2)</b>	108	208	4, 10	<b>1.2.1</b>	87	55	—	145.3/803.0	$<2.2 \times 10^{-16}$	0.819	1/14
I.2.1.1 (EAI2_MNL)	12	63	4, 10, 30–31					120.3/286.8	$3.64 \times 10^{-05}$	0.580	3/12
I.2.1.2 (EAI2_NTB)	94	20	4, 10, 17–35					106.4/272.8	$<2.2 \times 10^{-16}$	0.610	2/8
I.2.1.3	2	53	3–4, 10, 30–31					76.0/201.9	0.2207*	0.614	3/1
<b>I.2.2</b>	18	150	62 (not unique)	<b>1.2.2</b>	95	59	3470377 C/T	430.2/851.9	$3.22 \times 10^{-07}$	0.495	2/7

**Table 1.** Summary of findings for each sublineage in this study. The sublineage numbers in bold conforms to the nomenclature of Coll's *et al.*<sup>27</sup>, with the additional sublineages, named in this study, shown in regular fonts. Specific SNPs appeared in all members of the sublineages. All members of some sublineages shared DVR deletions, but the DVRs may have been deleted sporadically from members of other sublineages. Common specific SNPs in both studies indicates the number of specific SNPs that were the same in both Coll's and this study. The probability denotes the probability that the mean SNV distances of isolates in a group from the other members in the same group were different from the ones from other isolates that were not in the same group but in the same level of grouping, by the Wilcoxon rank-sum test.

several DVRs were not identified computationally but they were detected by experimental spoligotyping, shown in Table 2. The discrepancy in the other cases was only one or two DVRs.

The SNV phylogeny indicated that EAI2\_MNL, EAI2\_NTB and EAI6\_BGD1 were monophyletic clades, corresponding exactly to L1.2.1.1, L1.2.1.2, and L1.1.3 respectively. Their DR regions were characterized by multiple deletions compared with the EAI\_MRCA (Table 1). The multiple deletion events should result in low probability of homoplasy and indicate common ancestors.

Comparison between SNV phylogeny and spoligotypes confirmed the notion that the DR region evolved mostly by deletions. All L1.1.3 isolates shared the deletions of DVR33 and 56, while L1.1.3.1 had additional deletions of DVR6–7, 51 and 57. All L1.2.1 isolates shared the deletions of DVR4 and 10. Each of its three sublineages, L1.2.1.1, L1.2.1.2 and L1.2.1.3 had additional deletions of DVR30–31, DVR17–35 and DVR3 respectively.

An experimental EAI6\_BGD1 isolate did not belong to L1.1.3. In fact, the isolate still had the unspoligotyped DVR56. Extending the spoligotyping to cover the entire DR region would easily differentiate it from L1.1.3.

**Lack of correspondence between SNV phylogenetic trees and some spoligotypic clades.** There was insufficient information to confirm that EAI3 was monophyletic but isolates classified as EAI4\_VNM were found in both L1.2.2 and L1.1.1.1, suggesting that EAI4\_VNM might not be monophyletic. It was clear, however, that EAI1 and EAI5 were not monophyletic. Both belonged to many sublineages and appeared together in some sublineages.

Major sublineage	Sub-lineage	Num-ber	Number of isolates × known spoligotypes (SIT)	Number of isolates × unclassified spoligotypes
1.1.1	1.1.1.1	14	1 × 67777777413771 (342), 2 × 77773777413771 (618), 2 × 77777774413771 (139), 8 × 77777777413771 (236)	1 × 71777777003371
	1.1.1.2	26	1 × 73777777413771 (204), 3 × 77773777413731 (349), 1 × 77773777413771 (618), 2 × 7777777413731 (48), 1 × 7777777413671 (256), 11 × 7777777413771 (236), 1 × 777777773771 (100), 3 × 777777777771 (523)	1 × 73777770003771, 1 × 77771777413771 1 × 77777777411771*
	1.1.1.3	7	2 × 76377777413771 (792), 1 × 77777767413731 (1404), 3 × 7777777413771 (236)	1 × 77777777410071
	1.1.1.4	5	3 × 7777777413731(48), 1 × 77777777413771 (236)	1 × 77777767413371
	1.1.1.5	42	15 × 77417777413731 (493), 14 × 7741777741377 (1951), 3 × 774177757413771 orphan	1 × 57417777413731, 1 × 67416777413731, 3 × 77400017413771, 1 × 77407777410071, 1 × 77417777411771, 2 × 77417777413001, 1 × 77417777413701
	1.1.1.6	27	1 × 7777777413731(548)*, 26 × 77777777413771 (236)	None
	1.1.1.7	44	1 × 77771777413671 (orphan), 1 × 77757777413771(1372), 2 × 77777637413771 (orphan), 1 × 77777760000000 (786), 1 × 77777760000031 (773), 12 × 77777760000071 (1274), 4 × 7777777013771 (937), 1 × 7777777413071 (934), 3 × 7777777413700 (138), 11 × 7777777413771 (236)	4 × 70357777413771 1 × 77637777413771 2 × 77777677413700
	1.1.1.8	73	56 × 7777777413671(256)* (Discrepancy in 3 cases), 1 × 7577777413671 (orphan)*, 2 × 7777777413631 (947)	3 × 5777777413671, 1 × 7377777413671, 2 × 77777077413671, 2 × 77777607413671, 1 × 777777413411, 3 × 7777777413660, 1 × 7777777413661
	1.1.1.9	20	1 × 7177777413731 (1316), 2 × 6777777413731 (529), 1 × 7777677413731 (735), 1 × 7777777413131 (745), 6 × 7777777413631 (947), 1 × 7777777413730 (1801), 8 × 7777777413731 (48)	1 × 677760377413731
1.1.2	1.1.2.1	23	15 × 77577777413771 (939)	1 × 77500007413771, 1 × 77577757413771*, 1 × 7757777403171, 4 × 7757777413471, 1 × 7757777413761
	1.1.2.2	9	1 × 40003777413771 (8)*, 1 × 474000377413031 (1983), 1 × 4757777413771 (Orphan)*, 1 × 47700177413771 (1875), 1 × 4777777413031 (355), 2 × 4777777413071 (11), 1 × 4777777413731 (1182)*	1 × 47577777413051*
1.1.3 (EAI6_BGD1)	1.1.3.1	21	13 × 77777757413371 (292), 1 × 77777700003371 (1391), 2 × 77777743413371 (Orphan), 1 × 77777747413371 (1390)	1 × 77757775413371 1 × 777767700003371* 1 × 77777707413371 1 × 77777750403371
	1.1.3.2	17	12 × 77777757413771 (591), 2 × 73777757413771 (orphan)	3 × 77777657413771
	1.1.3.3	15	11 × 77777757413771 (591)	1 × 777701757413771, 1 × 77777703413771, 2 × 77777757412771*
1.2.1 (EAI2)	1.2.1.1 (EAI2_MNL)	12	1 × 677767477413771 (1490), 1 × 67777477413701 (483), 1 × 67777477413751 (287), 9 × 67777477413771 (19)	None
	1.2.1.2 (EAI2_NTB)	94	83 × 67400003413771 (89)* (Discrepancy in 6 cases) 1 × 07400003413771 (orphan), 1 × 60000000000000 (orphan), 4 × 67400002000071 (orphan)	1 × 47400003413771, 2 × 67400003413700, 2 × 67400003413711
	1.2.1.3	2	1 × 47777477413771 (413)	1 × 47777477413731
1.2.2	1.2.2	18	2 × 5777777413731 (477), 1 × 77776777413731 (1251)*, 2 × 7777774413731 (514), 1 × 7777777413631 (947), 1 × 7777777413711 (517), 6 × 7777777413731 (48)	1 × 77617775413731, 1 × 7777774403731*, 1 × 7777774412731, 1 × 7777775413731, 1 × 7777777411631
	Un-classified	11	2 × 7377777413771 (204), 3 × 7777774000071 (944), 1 × 77777757413771 (591), 2 × 7777777413711 (517), 1 × 7777777413731 (48)	1 × 77363777413771, 1 × 77763777413411

**Table 2.** The number of isolates with various experimental spoligotypes identified in each sublineage, listed in ascending order of the octal codes. The spoligotypes that are found in more than one sublineage are indicated in bold typeface. No unclassified spoligotypes appeared in two sublineages. (SIT denotes the spoligotype international type). \*The spoligotypes, of which some isolates had different predicted spoligotypes.

There were 60 isolates, with SIT236 (7777777413771), typical for EAI5, belonging to 6 sublineages of L1.1.1 (Table 2). The isolates contained all spoligotyped DVRs, except DVR39–42 and 44. Thus, SIT236 was not defined by any specific deletion but instead by the lack of any deletion in the spoligotyped DVRs. Twenty-one isolates

(34%) contained all 62 DVRs apart from the six EAI-defining DVRs; therefore, they should have the DR region similar to the EAI-MRCA. This result indicated that some descendants of EAI-MRCA had maintained the DR region without further deletions. Nevertheless, they separately accumulated different mutations and became different sublineages of L1.1.1; specifically, L1.1.1.1- L1.1.1.4 and L1.1.1.7.

A similar situation was observed with EAI1. The common EAI1 spoligotype, SIT48 (77777777413731) has only one more deleted DVR, DVR62, compared with SIT236. In this study, 14 of 20 isolates with SIT48, belonging to L1.2.2, L1.1.1.4 and L1.1.1.9, retained all the other unspoligotyped DVRs. As L1.2.2 separated from the other sublineages early in the evolution of L1 (Fig. 1), its DVR62 deletion must also occur very early. However, because DVR62-deleted isolates were also observed in many sublineages of L1.1, sometimes in the same sublineage as isolates with the intact DVR62, the DVR62 deletions in L1.1 must be separate, homoplastic, events. Isolates with SIT48 and SIT236 have been reported in wide geographic areas (Supplementary Table S7). Still, it cannot be concluded that the isolates belonged to the same sublineages.

There were some sublineages associated with single DVR deletions such as L1.1.2.1-DVR18, L1.1.1.5-DVR18-21, L1.1.1.8-DVR53, and L1.1.1.6-DVR57-60. However, the associations were not completely specific and the single deletion event might allow for homoplasmy, as in the case of DVR62 deletion.

**Limited association between SNP phylogeny and IS6110 copy number.** *M. tuberculosis* isolates from South and Southeast Asia usually have a single copy or low copy numbers of IS6110<sup>32</sup>. Our study revealed that the isolates could belong to many sublineages except for L1.2.1 and L1.1.3 (Supplementary Table S3). Unless the DVR34 and DVR35 were deleted, the single copy was always found between them.

**No association of patient profiles with sublineages except for age.** The profiles of the patients infected by each major sublineage are shown in Table 3, of whom 89% were Thais, 3% were foreigners and the remaining 8% belonged to one of the several hill tribes. The majority of patients (73.8%) were male with a median age of 50 years. The proportions of patients over age 49 infected by various sublineages were significantly different (Chi-square test,  $p = 0.0246$ ). The relative risk of patients over age 49 being infected by L1.1.1 compared to the patients with younger age was 1.29 ( $p = 0.0019$ , CI:1.099-1.522). There was no association between HIV infection and the sublineages.

**Similarity of population structures between various countries.** As the informations regarding the SNP-based genetic population structures of L1 in various endemic countries are very limited while the informations on spoligotypes are widely available, we investigated the similarity of population structures of L1 by examining the correlation of spoligotype profiles and the similarity of the most predominant clades. The SNV phylogeny above allowed the exclusion of the homoplastic spoligotypes, which belonged to many sublineages, such as SIT48 and SIT236.

**Correlation between spoligotypes in various countries.** The correlation of frequencies of EAI spoligotypes between various studies was calculated by excluding the data for homoplastic spoligotypes. Although each study had different and sometimes unspecified sampling designs, which mandates caution in interpreting the results, several general patterns can be observed (Supplementary Table S8). There were no correlations of spoligotype profiles between countries in East Africa (excluding Madagascar) and Southeast Asia.

Strong correlations ( $>0.8$ ) were usually found between studies performed in the same country, as expected, which might not be true if a large number of samples were intentionally selected from patients with specific forms of the disease, such as extrapulmonary or drug-resistant tuberculosis. There were numerous studies in India, where the percentages of L1 are known to be high in the southern and low in the northern part of the country. Nevertheless, the L1 spoligotype profiles were generally similar throughout the country, excluding eastern India.

Correlations between studies in bordering countries were usually intermediate and higher than correlations between countries without common borders. Strong correlations were identified between Pakistan and Indian studies, particularly when homoplastic spoligotypes were excluded, conforming to the close historical relationship between the two countries. There were also strong correlations between the spoligotype profiles in Bangladesh and Myanmar (Supplementary Table S8).

Intriguing patterns of spoligotype profile correlations were observed in Southeast Asian countries. Countries in islands of Southeast Asia (ISEA), Indonesia and the Philippines, as well as Singapore and Taiwan are separated by seas but share similar L1 spoligotype profiles comprising predominantly EAI2\_MNL (L1.2.1.1). In contrast, there was low similarity of spoligotype profiles between MSEA countries. In all cases, there were some levels of similarity of spoligotype profiles only between countries that shared common borders. The correlation coefficients between spoligotype profiles between countries without common borders were always low (Supplementary Table S8).

**Literature review indicating that the predominant sublineages in each country were usually different.** The reported predominant spoligotypes in various countries were usually different, as shown in Fig. 3 and Supplementary Table S7. Countries that had similar profiles usually shared the same predominant spoligotypes.

**Predominant sublineages in Chiangrai were not reported frequently in neighbouring countries.** The two most common sublineages (L1.2.1.2 and L1.1.1.8) contributed to a third of L1 isolates in Chiangrai. Of the 94 members of L1.2.1.2, 83 (88%) had SIT89 or EAI2\_NTB while the remaining 11 had new spoligotypes that could be derived from SIT89 by additional DVR deletions. The family was initially recognized by IS6110-RFLP in Nonhaburi, central Thailand<sup>32</sup>. All 73 L1.1.1.8 isolates had a characteristic deletion of DVR53, with 56 (77%)

	L1.1.1	L1.1.2	L1.1.3	L1.2.1	L1.2.2	All
Total	269	32	53	108	18	480
<b>Sex</b>						
Male	198	27	37	78	14	354
Female	71	5	16	30	4	126
<b>Ages</b>						
<19	1	0	2	4	0	7
20–29	24	2	4	12	1	43
30–39	42	6	15	25	4	92
40–49	47	8	12	20	5	92
50–59	57	5	7	18	1	88
60–69	50	9	10	11	3	83
>70	48	2	3	18	4	75
Average ages	53.2	50.3	45.6	48.5	51.9	50.7
Median ages	54	49	43	46	47	50
% with ages >49*	57.6	50.0	37.7	43.5	44.4	51.3
<b>Ethnicity</b>						
Thai	238	26	45	103	17	429
Hill tribe	24	3	6	3	1	37
Foreigner	7	3	2	2	0	14
<b>Clinical presentation</b>						
Pulmonary	258	32	49	102	17	458
Pulmonary and Extrapulmonary	9	0	4	6	1	20
Extrapulmonary	2	0	0	0	0	2
<b>HIV status</b>						
Positive	49	5	12	29	4	99
Negative	215	27	40	76	13	371
Unknown	5	0	1	3	1	10

**Table 3.** Demographic and clinical profiles of patients infected by different sublineages in this study. \*The proportions of the patients over age 49 were different between various sublineages (Chi-square test,  $p = 0.0246$ ). The relative risk of infecting by L1.1.1 of patients over age 49, compared to the ones being younger was 1.29 ( $p = 0.0019$ , 95% CI: 1.099–1.522).

isolates having SIT256, while the others had spoligotypes that could be derived from SIT256 by deletion. Both L1.2.1.2 and L1.1.1.8 had very low intragroup MPSDs of 106.4 and 174.4 and very high fixation indices of 0.61 and 0.63 respectively, indicating a close genetic relationship within each group. Together with the high prevalence, these results suggested high transmission activities of both sublineages.

In countries bordering Thailand, Cambodia, Malaysia and Myanmar, both SIT89 and SIT256 were reported but at much lower frequencies (Supplementary Table S7). This result suggested that their transmission activities were greater in Chiangrai. It can, therefore, be hypothesized that L1.2.1.2 and L1.1.1.8 might be better adapted to the populations in Chiangrai.

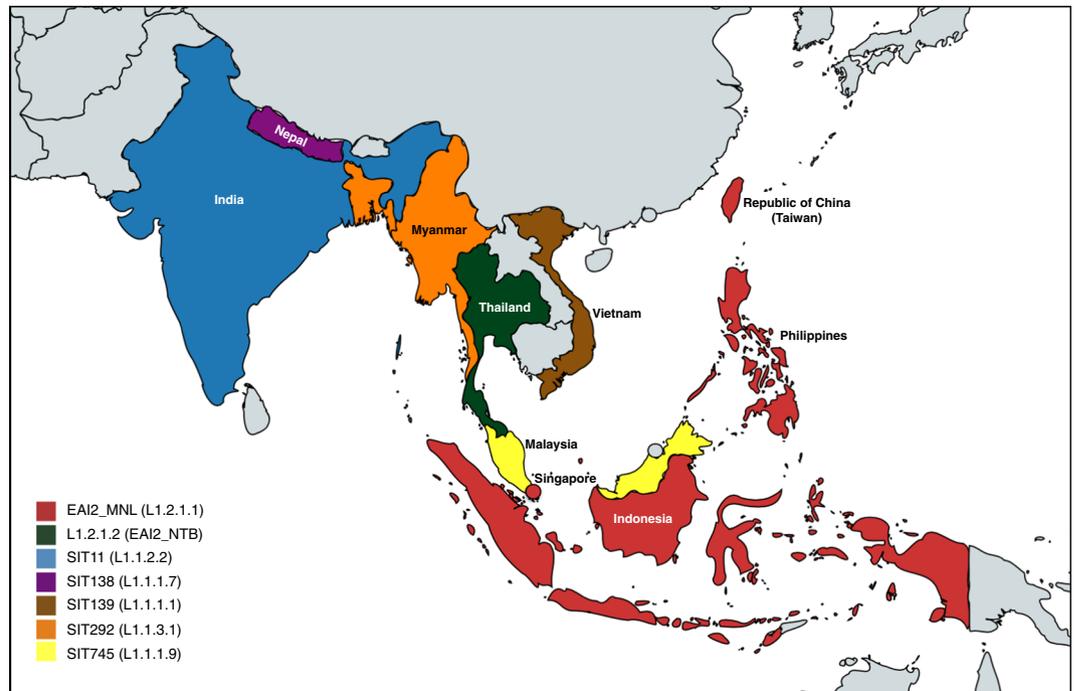
SIT292, the most common spoligotype in L1.1.3.1, was predominant in Myanmar and Bangladesh and also found in many patients in Chiangrai. People from both countries contributed substantially to the workforce population in Chiangrai (Supplementary Table S7).

## Discussion

In this study, we analysed WGS of 480 L1 isolates and refined its classification. The results should facilitate comparison of WGS information across countries. Comparisons of phylogenetic trees with spoligotypes indicated the benefits of using all 68 DVRs instead of only 43 for spoligotyping.

The availability of spoligotype information in many countries indicated that the predominant sublineages in each country varied with geography and apparently with ethnicity.

EAI2\_MNL or L1.2.1.1 was highly prevalent in the Philippines, accounting for more than 80% of all *M. tuberculosis* isolates, and in Southern Taiwan and Indonesia, where it accounted for >90% and approximately 40% of EAI isolates respectively (Supplementary Tables S6 and S7). ISEA, including the Philippines and Indonesia, are home to people who speak Austronesian languages that originated in Southern Taiwan approximately 5000 years ago and spread throughout the Pacific islands and to Madagascar<sup>33,34</sup>. A few spoligotyping studies in other island countries, such as in Kiribati<sup>35</sup> and Papua New Guinea<sup>36,37</sup>, revealed the rarity of the EAI isolates. The rare EAI isolates, however, usually belonged to EAI2\_MNL, which was also reported from Hawaii and Guam<sup>38</sup>. Thus, L1.2.1.1 may associate with Austronesians either due to the founder effect or co-evolution. Furthermore, the immediate prediction is that the common spoligotype in Madagascar, SIT109 (EAI8\_MDG)<sup>39</sup>, should also belong to L1.2.1.1.



**Figure 3.** Area map showing the predominant sublineages of L1 in different countries in Southeast Asia, deduced from their spoligotype reports. There were no obvious predominant strains in Cambodia and there was no information from Lao.

In contrast to ISEA, countries in MSEA are geographically closer but populated by ethnic groups that speak different language families, suggesting different genetic roots. Most Vietnamese people belong to the Kinh ethnic group and speak the Vietic branch of Austroasiatic language<sup>40</sup>, while Cambodians speak the Khmer branch. The Thai language belongs to the Tai-Kadai family while Burmese belongs to the Tibeto-Burman family. The Malay language belongs to the Malayo-Polynesian branch of the Austronesian family. Each country also had a different predominant spoligotype belonging to different sublineages (Fig. 3), while there was no clear predominant spoligotype in Cambodia (Supplementary Table S7).

The differences in the predominant L1 strains and spoligotype profiles in MSEA countries are intriguing because they have a history of extensive interactions accompanied by alternative expansion and contraction of territorial control for over a millennium. Nevertheless, the inter-country variations of *M. tuberculosis* were still obvious, suggesting a contribution of co-adaptation between the bacteria and their human hosts. The results herein, therefore, provided supportive evidence for the co-adaptation hypothesis in the case of L1, similar to the finding in lineage 4<sup>7</sup>. Due to the poor uniformity of each spoligotyping study, as shown in supplementary Table S6, rigorous confirmation of the hypothesis will require WGS of representative samples of bacterial populations from many countries around the Indian Ocean.

The predominant L1 spoligotype in Malaysia was SIT745 (27%), while EAI2\_MNL accounted for approximately 9%<sup>16</sup>. Although the native language of many Malaysians is Austronesian, similar to people in ISEA, current Malaysians may be a genetic admixture between people who were originally in MSEA and ISEA<sup>41</sup>.

The most common spoligotype in Vietnam was uniquely SIT139, a rare spoligotype classified as L1.1.1.1 in this study, accounting for 48% of Vietnamese L1 isolates<sup>17</sup>, which might indicate a specific adaptation. In contrast, L1.2.1.1 was rare in Vietnam. This result is intriguing because the Cham people in the historical Champa Kingdom, in what is now southern Vietnam, spoke Austronesian. The kingdom lasted for several centuries and afforded ample interactions between Cham and Kinh, the major ethnic group in Vietnam. The rarity of L1.2.1.1 could be a consequence of the Champa Kingdom acquiring mainly Austronesian languages and culture rather than a large migration of genetic Austronesians<sup>42</sup>.

Existing studies on the transmissibility of L1 supported the co-evolution hypothesis. A population-based study in Southern Taiwan revealed that the genotypic clustering rates of 60.8% among EAI strains<sup>43</sup>, which mostly belonged to L1.2.1.1, suggesting high transmissibility. Another study in Vietnam reported the clustering rate among EAI2\_MNL (L1.2.1.1), an uncommon sublineage in that region, of only 18.3% while the clustering rate among EAI4\_VNM, the predominant clade including SIT139, was 77.4%<sup>44</sup>. This result suggested that L1.2.1.1 and SIT139 spread differently in both populations, conforming to the co-adaptation hypothesis.

Robust phylogenetics provides a foundation for genotype-phenotype correlation studies. In general, it is difficult to compare phenotypes of *M. tuberculosis* lineages at the population level across several studies. Although many recent studies provided genotypic details of the studied isolates, their phylogenetic relationships were usually not considered in the analysis<sup>45</sup>. Moreover, many studies used the phenotypes of the other strains in the same studied areas as the controls. The controls, therefore, varied between studies in both proportions of lineages and sublineages. Moreover, in most studies, information regarding host population genetic structures are usually not available.

*In vitro* phenotypic studies of *M. tuberculosis* typically involved a small number of strains and rarely had sub-lineage information, rendering comparison between studies difficult. A few recent studies, involving L1 isolates with sublineage information, included the autophagy study of EAI2\_NTB (L1.2.1.2)<sup>46</sup>. A number of EAI isolates belonging to SIT1390 (L1.1.3.1)<sup>23</sup> or EAI2\_MNL (L1.2.1.1)<sup>22,23,47</sup> were shown to induce high levels of inflammatory cytokines. Unfortunately, there are no known *in vitro* studies involving L1.1.1 or L1.1.2 isolates that merit further exploration.

India harbours a quarter of the world patient population<sup>1</sup>, its predominant strain, SIT11 belonging to L1.1.2.2 (Supplementary Table S7), may be one of the genotypes infecting the highest absolute number of people, similar to L1.2.1.1, which is common in ISEA<sup>1</sup>. Recently, it was found that 90% of *M. tuberculosis* in Tamil Nadu, Southern India, where the Chingleput BCG trials had been conducted<sup>26</sup>, were either EAI3, including SIT11, or EAI5<sup>48</sup>, which should belong to L1.1.2 or L1.1.1 respectively. Further genomics and pathogenesis studies of the sublineages may provide important clues for the development of new tuberculosis vaccines.

The co-adaptation between *M. tuberculosis* and its hosts may explain the difficulty in consistently identifying human genetic markers associated with tuberculosis<sup>49</sup>. Considering that the host genetic susceptibility risks for tuberculosis are likely oligogenic and may not be similar across multiple populations with large genetic distances, it is possible that host susceptibility follows a genetic heterogeneity model, in which a group of people with specific genetic susceptibility allows better local spreading of some sublineages of *M. tuberculosis*<sup>50</sup>. The susceptibility may be mild and significant only among those who have high genetic similarity, such as family members, which may result in a higher prevalence in a genetically related population. Consequently considering the genotypic profiles of *M. tuberculosis* as strata of differential susceptibility risk factors in host susceptibility analysis may result in the successful identification of the host genetic factors<sup>9,10</sup>.

In conclusion, 18 tentative L1 sublineages belonging to three major groups were identified. The limitations of standard spoligotyping were revealed, favouring extended spoligotyping methods using all 68 DVRs. The phylogenetic trees suggested relatively high transmission activities in some sublineages, inferring possible selective advantages. However, some common spoligotypes in neighbouring countries were uncommon in Chiangrai, suggesting different levels of transmission in different countries. This conforms to the notion that *M. tuberculosis* sublineage adapt to the host populations. This finding has implications for studies on the host-pathogen relationship such as GWAS, and may have implications in the urgent need for vaccine testing and development.

## Materials and Methods

**Settings.** This study included *M. tuberculosis* isolates from patients in Chiangrai Province, Northern Thailand from 2003–2010. Chiangrai borders both Myanmar and Lao and is an important transportation hub with China through the Mekong River. The population of the province consists of approximately 1.2 million people with a tuberculosis incidence rate of 152.6/100,000 population in 2011.

The isolates were collected as part of a tuberculosis cohort study initiated by the Japan Research Institute of Tuberculosis (RIT), the Japan Anti-Tuberculosis Association (JATA), and the Ministry of Public Health, Thailand. The study recruited all newly-diagnosed culture positive tuberculosis patients with all nationality who were also willing to participate in a human genetic study from all public hospitals in Chiangrai. During the studied period, there were 15,805 new cases in Public Hospital Tuberculosis Registration System. Bacterial isolates were obtained from 7148 patients, of which both bacterial and host DNA were obtained from 1187 patients. 1174 *M. tuberculosis* isolates were successfully sequenced and 480 were identified to be Indo-Oceanic strains.

**Ethics Statement.** The project was approved by the Ethical Committees of Chiangrai Prachanukroh Hospital, Chiangrai and the Thai Ministry of Public Health. Informed consent was obtained from all participants and/or their legal guardians. All methods were performed in accordance with the relevant guidelines and regulations.

**Bacteria.** Bacterial samples from 1187 patients were successfully regrown in Lowenstein-Jensen medium in an appropriately contained clinical microbiology laboratory in Chiangrai using standard biosafety protocols and equipment. The bacteria were heat-killed and DNA was prepared as previously described<sup>32</sup>. All the processes were performed in Class II biosafety cabinets.

**Bacterial genotyping.** *Large sequence polymorphisms (LSP).* Lineages of the isolates in this study were initially identified by LSP using PCR primers specific to TbD1 (for differentiation between ancestral and modern strains), RD239 (for lineage 1), RD105 (for lineage 2), RD750 (for lineage 3) and 7 bp sequence at pks15/1 (for lineage 4)<sup>51</sup>. The primer sequences are shown in Supplementary Table S9.

The 487 isolates were identified as Indo-Oceanic strains due the presence of the TbD1 and the absence of RD239.

**Spoligotyping.** Spoligotyping was performed using a commercial kit (Ocimum Biosolutions, India) as previously described<sup>52</sup>. Spoligotyping investigates the presence or absence of 43 of the 68 DVRs. The results were coded into a binary pattern of 43 digits, which is typically converted to an octal code by the successive combination of three binary digits from the left, leaving the 15<sup>th</sup> code as still binary. *M. tuberculosis* L1 typically has EAI spoligotypes, i.e. the 10–12th octal code of 413. Some octal codes have been designated as spoligotype international type (SIT) and assigned to one of the various clades, such as EAI1, EAI2\_MNL, and EAI2\_NTB, among others. The SIT and clade assignment was performed by querying SITVITWEB ([http://www.pasteur-guadeloupe.fr:8081/SITVIT\\_ONLINE/](http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE/); updated February 17, 2017)<sup>30</sup>. Spoligotypes were also predicted from whole-genome sequences using *SpolPred* software<sup>53</sup>.

**Whole-genome sequencing and SNV analysis.** *M. tuberculosis* samples were sequenced on the Illumina HiSeq. 2000 platform at the Wellcome Trust Sanger Institute, UK. The system produced paired-end reads in FastQ file format. We used Trimmomatic version 0.36<sup>54</sup> to collect only the paired reads. Five isolates were not further analysed because the numbers of available reads were too small. The raw reads were filtered and discarded if their phred quality scores were lower than 20. The remaining reads were aligned by the BWA program<sup>55</sup> and SAMtools<sup>56</sup> using *M. tuberculosis* H37Rv (GenBank NC\_000962.3) as the reference genome with alignment score more than 50, SAMtools base quality more than 23 and read depth more than 10. Two samples appeared to contain mixed nucleotide sequences and were also not further studied. The SNVs that were present in any drug-resistance gene, mobile genetic element, phage, PE/PPE region and non-homozygous SNVs were discarded. The remaining SNVs had been converted to an SNV-supermatrix using an in-house Python script before being used in the phylogenetic analysis. The sequencing data for the 480 samples used in this paper were submitted to the European Nucleotide Archive (ENA) of EMBL-EBI mirrored in the Sequence Read Archive (SRA) database. Actual read sequences can be queried and downloaded directly from the SRA database using the accession numbers listed in Supplementary Table S10.

**Phylogenetics analysis.** SNVs of 480 Indo-Oceanic isolates were used for phylogenetic tree construction, with 24 isolates from the lineage 2–4 as outgroups. We analysed the generated supermatrix with two phylogenetic analytic methods, maximum likelihood (ML) and Bayesian inference (BI).

For ML analysis, we used PhyML<sup>57</sup> via the SeaView program<sup>58</sup>. The core tree (starting tree) was generated using the BioNJ method. The analysis was run using GTR model, using the nearest neighbor interchange (NNI) method. The bootstrap analysis was calculated for branch support with 1,000 replications of pseudo-data.

For the BI methods, we used the MrBayes program<sup>54</sup>. The parameter “4by4” was used as the nucleotide substitution model. The BI analysis was performed with two simultaneous runs with four n-chain for each run. The analysis was run for approximately 10–20 million generations. The analysis was terminated when the average standard deviation of split frequencies was lower than 0.01. The first 25% of sample trees were discarded as burned-in. All trees were visualized using the FigTree program version 1.4.2. (<http://tree.bio.ed.ac.uk/software/figtree/>).

Principal Component Analysis was performed using Jalview 2.8.2<sup>57</sup>. The analysis was conducted at multiple levels of sublineage classification. The plots between the first three eigenvectors were examined.

Pairwise SNV distances were calculated using MEGA5<sup>56</sup>. The fixation index ( $F_{ST}$ ) was calculated based on the following formula  $(\pi_{\text{between group}} - \pi_{\text{within group}}) / \pi_{\text{between group}}$ , where  $\pi_{\text{within group}}$  is the average pairwise SNV distance within a group and  $\pi_{\text{between group}}$  is the average pairwise SNV distance between all members in the group and all members not in the same group but in the same level of grouping. The statistical tests for the difference between the within-group averages and between-group averages were performed using Wilcoxon rank-sum test at  $\alpha = 0.05$  with Bonferroni correction multiple testing using R statistical package version 3.3.2.

**Analysis of the Direct Repeat (DR) region and IS6110 copy number.** The sequence contigs of each isolate were constructed *de novo* using velvet software<sup>59</sup>. Analysis of the DR region in each genome was performed by carrying out a blastn search of the contig files containing the DR region. The identified contigs were mapped to areas around the position of 3119181–3123573 in the genome of H37Rv (GenBank NC\_000962.3). The presence or absence of each of the 68 known direct variable region (DVR) sequences was determined and recorded as shown in Supplementary Table S5.

The number of IS6110 in the genome of each isolate was identified by ISMapper using the provided IS6110 sequence<sup>60</sup> as the query and H37Rv as the reference genome with default parameters.

**Analysis of phenotypic data.** The patient characteristics were described in descriptive statistics (as presented in Table 3). The association between phenotypes and the five major sublineages of L1 were evaluated by one-tailed Pearson Chi-square test. The P-values less than 0.05 were considered significant. Association with ages were analysed by categorizing patients into two groups, with the age over 49 and 49 or less. The relative risk of patients over age 49 for being infected by L1.1.1 strains was calculated compared to the younger patients.

**Comparison of spoligotypes between countries.** Relevant research articles were identified by PubMed search using the keywords, spoligotype or genotype and tuberculosis and names of the countries in the ASEAN, South Asian and Pacific islands. The full papers were retrieved and examined for the presence of detailed descriptions of spoligotypes. Forty-one articles were identified with 6 having less than 20 EAI isolates in each study. The latter group was excluded. The list of retrieved papers is shown in Supplementary Table S6. The numbers of EAI isolates and isolates belonging to each reported SIT were tabulated, as shown in Supplementary Table S7, with the frequency of each spoligotype calculated. The data for spoligotypes that could appear in more than one sublineages, SIT48, 204, 236, 517, 618 and 947, were excluded. Then pairwise correlations between the frequencies were evaluated by calculating Pearson correlation coefficients. A higher correlation coefficient indicated a greater similarity of the spoligotype profiles between two studies.

## References

1. *Global tuberculosis report 2017*. (World Health Organization, 2017).
2. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin Immunol* **26**, 431–444, <https://doi.org/10.1016/j.smim.2014.09.012> (2014).
3. Hershberg, R. Human host range of *Mycobacterium tuberculosis*. *Nat Genet* **48**, 1453–1454, <https://doi.org/10.1038/ng.3724> (2016).

4. Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W. & Small, P. M. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci USA* **101**, 4871–4876, <https://doi.org/10.1073/pnas.0305627101> (2004).
5. Gagneux, S. *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* **103**, 2869–2873, <https://doi.org/10.1073/pnas.0511240103> (2006).
6. Fenner, L. *et al.* HIV infection disrupts the sympatric host-pathogen relationship in human tuberculosis. *PLoS Genet* **9**, e1003318, <https://doi.org/10.1371/journal.pgen.1003318> (2013).
7. Stucki, D. *et al.* *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet* **48**, 1535–1543, <https://doi.org/10.1038/ng.3704> (2016).
8. van Crevel, R. *et al.* Infection with *Mycobacterium tuberculosis* Beijing genotype strains is associated with polymorphisms in SLC11A1/NRAMP1 in Indonesian patients with tuberculosis. *J Infect Dis* **200**, 1671–1674, <https://doi.org/10.1086/648477> (2009).
9. Toyo-Oka, L. *et al.* Strain-based HLA association analysis identified HLA-DRB1\*09:01 associated with modern strain tuberculosis. *HLA*. <https://doi.org/10.1111/tan.13070> (2017).
10. Omae, Y. *et al.* Pathogen lineage-based genome-wide association study identified CD53 as susceptible locus in tuberculosis. *J Hum Genet*. <https://doi.org/10.1038/jhg.2017.82> (2017).
11. Brites, D. & Gagneux, S. Co-evolution of *Mycobacterium tuberculosis* and *Homo sapiens*. *Immunol Rev* **264**, 6–24, <https://doi.org/10.1111/imr.12264> (2015).
12. Niemann, S., Merker, M., Kohl, T. A. & Supply, P. Impact of genetic diversity on the biology of *Mycobacterium tuberculosis* complex strains *Microbiology Spectrum* **4**, TBTB2-0022-2016 (2016).
13. Tientcheu, L. D. *et al.* Immunological consequences of strain variation within the *Mycobacterium tuberculosis* complex. *Eur J Immunol* **47**, 432–445, <https://doi.org/10.1002/eji.201646562> (2017).
14. Tsolaki, A. G. *et al.* Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci USA* **101**, 4865–4870, <https://doi.org/10.1073/pnas.0305634101> (2004).
15. Mbugi, E. V. *et al.* Mapping of *Mycobacterium tuberculosis* Complex Genetic Diversity Profiles in Tanzania and Other African Countries. *PLoS One* **11**, e0154571, <https://doi.org/10.1371/journal.pone.0154571> (2016).
16. Ismail, F. *et al.* Study of *Mycobacterium tuberculosis* complex genotypic diversity in Malaysia reveals a predominance of ancestral East-African-Indian lineage with a Malaysia-specific signature. *PLoS One* **9**, e114832, <https://doi.org/10.1371/journal.pone.0114832> (2014).
17. Nguyen, V. A. *et al.* High prevalence of Beijing and EAI4-VNM genotypes among *M. tuberculosis* isolates in northern Vietnam: sampling effect, rural and urban disparities. *PLoS One* **7**, e45553, <https://doi.org/10.1371/journal.pone.0045553> (2012).
18. Click, E. S., Winston, C. A., Oeltmann, J. E., Moonan, P. K. & Mac Kenzie, W. R. Association between *Mycobacterium tuberculosis* lineage and time to sputum culture conversion. *Int J Tuberc Lung Dis* **17**, 878–884, <https://doi.org/10.5588/ijtld.12.0732> (2013).
19. Buu, T. N. *et al.* The Beijing genotype is associated with young age and multidrug-resistant tuberculosis in rural Vietnam. *Int J Tuberc Lung Dis* **13**, 900–906 (2009).
20. Shanmugam, S., Selvakumar, N. & Narayanan, S. Drug resistance among different genotypes of *Mycobacterium tuberculosis* isolated from patients from Tiruvallur, South India. *Infect Genet Evol* **11**, 980–986, <https://doi.org/10.1016/j.meegid.2011.03.011> (2011).
21. Click, E. S., Moonan, P. K., Winston, C. A., Cowan, L. S. & Oeltmann, J. E. Relationship between *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis. *Clin Infect Dis* **54**, 211–219, <https://doi.org/10.1093/cid/cir788> (2012).
22. Portevin, D., Gagneux, S., Comas, I. & Young, D. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS Pathog* **7**, e1001307, <https://doi.org/10.1371/journal.ppat.1001307> (2011).
23. Reiling, N. *et al.* Clade-specific virulence patterns of *Mycobacterium tuberculosis* complex strains in human primary macrophages and aerogenically infected mice. *MBio* **4**, <https://doi.org/10.1128/mBio.00250-13> (2013).
24. Uplekar, M. *et al.* WHO's new end TB strategy. *Lancet* **385**, 1799–1801, [https://doi.org/10.1016/S0140-6736\(15\)60570-0](https://doi.org/10.1016/S0140-6736(15)60570-0) (2015).
25. Collins, C. H., Yates, M. D. & Grange, J. M. Subdivision of *Mycobacterium tuberculosis* into five variants for epidemiological purposes: methods and nomenclature. *J Hyg (Lond)* **89**, 235–242 (1982).
26. Trial of BCG vaccines in south India for tuberculosis prevention: first report—Tuberculosis Prevention Trial. *Bull World Health Organ* **57**, 819–827 (1979).
27. Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* **5**, 4812, <https://doi.org/10.1038/ncomms5812> (2014).
28. Vongpaisarnsin, K., Listman, J. B., Malison, R. T. & Gelernter, J. Ancestry informative markers for distinguishing between Thai populations based on genome-wide association datasets. *Leg Med (Tokyo)* **17**, 245–250, <https://doi.org/10.1016/j.legalmed.2015.02.004> (2015).
29. Brunelli, A. *et al.* Y chromosomal evidence on the origin of northern Thai people. *PLoS One* **12**, e0181935, <https://doi.org/10.1371/journal.pone.0181935> (2017).
30. Demay, C. *et al.* SITVITWEB—a publicly available international multimer database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infect Genet Evol* **12**, 755–766, <https://doi.org/10.1016/j.meegid.2012.02.004> (2012).
31. Flores, L. *et al.* Large sequence polymorphisms classify *Mycobacterium tuberculosis* strains with ancestral spoligotyping patterns. *J Clin Microbiol* **45**, 3393–3395, <https://doi.org/10.1128/JCM.00828-07> (2007).
32. Palittapongarnpim, P. *et al.* Restriction fragment length polymorphism study of *Mycobacterium tuberculosis* in Thailand using IS6110 as probe. *Int J Tuberc Lung Dis* **1**, 370–376 (1997).
33. Soares, P. A. *et al.* Resolving the ancestry of Austronesian-speaking populations. *Hum Genet* **135**, 309–326, <https://doi.org/10.1007/s00439-015-1620-z> (2016).
34. Regueiro, M. *et al.* Austronesian genetic signature in East African Madagascar and Polynesia. *J Hum Genet* **53**, 106–120, <https://doi.org/10.1007/s10038-007-0224-4> (2008).
35. Aleksic, E. *et al.* First molecular epidemiology study of *Mycobacterium tuberculosis* in Kiribati. *PLoS One* **8**, e55423, <https://doi.org/10.1371/journal.pone.0055423> (2013).
36. Ballif, M. *et al.* Genetic diversity of *Mycobacterium tuberculosis* in Madang, Papua New Guinea. *Int J Tuberc Lung Dis* **16**, 1100–1107, <https://doi.org/10.5588/ijtld.11.0779> (2012).
37. Ley, S. D. *et al.* Diversity of *Mycobacterium tuberculosis* and drug resistance in different provinces of Papua New Guinea. *BMC Microbiol* **14**, 307, <https://doi.org/10.1186/s12866-014-0307-2> (2014).
38. Frink, S. *et al.* Rapid deletion-based subtyping system for the Manila family of *Mycobacterium tuberculosis*. *J Clin Microbiol* **49**, 1951–1955, <https://doi.org/10.1128/JCM.01338-10> (2011).
39. Ferdinand, S. *et al.* A study of spoligotyping-defined *Mycobacterium tuberculosis* clades in relation to the origin of peopling and the demographic history in Madagascar. *Infect Genet Evol* **5**, 340–348, <https://doi.org/10.1016/j.meegid.2004.10.002> (2005).
40. Miranda-Barros, F. *et al.* Y Chromosome STR haplotypes in different ethnic groups of Vietnam. *Forensic Sci Int Genet* **22**, e18–e20, <https://doi.org/10.1016/j.fsigen.2016.02.007> (2016).
41. Norhalifah, H. K., Syaza, F. H., Chambers, G. K. & Edinur, H. A. The genetic history of Peninsular Malaysia. *Gene* **586**, 129–135, <https://doi.org/10.1016/j.gene.2016.04.008> (2016).
42. Peng, M. S. *et al.* Tracing the Austronesian footprint in Mainland Southeast Asia: a perspective from mitochondrial DNA. *Mol Biol Evol* **27**, 2417–2430, <https://doi.org/10.1093/molbev/msq131> (2010).

43. Chen, Y. Y. *et al.* Molecular epidemiology of tuberculosis in Kaohsiung City located at southern Taiwan, 2000–2008. *PLoS One* **10**, e0117061, <https://doi.org/10.1371/journal.pone.0117061> (2015).
44. Buu, T. N. *et al.* Increased transmission of Mycobacterium tuberculosis Beijing genotype strains associated with resistance to streptomycin: a population-based study. *PLoS One* **7**, e42323, <https://doi.org/10.1371/journal.pone.0042323> (2012).
45. Coscolla, M. & Gagneux, S. Does M. tuberculosis genomic diversity explain disease diversity? *Drug Discov Today Dis Mech* **7**, e43–e59, <https://doi.org/10.1016/j.ddmec.2010.09.004> (2010).
46. Haque, M. F. *et al.* Resistance to cellular autophagy by Mycobacterium tuberculosis Beijing strains. *Innate Immun* **21**, 746–758, <https://doi.org/10.1177/1753425915594245> (2015).
47. Chen, Y. Y. *et al.* The pattern of cytokine production *in vitro* induced by ancient and modern Beijing Mycobacterium tuberculosis strains. *PLoS One* **9**, e94296, <https://doi.org/10.1371/journal.pone.0094296> (2014).
48. Narayanan, S. *et al.* Genomic interrogation of ancestral Mycobacterium tuberculosis from south India. *Infect Genet Evol* **8**, 474–483, <https://doi.org/10.1016/j.meegid.2007.09.007> (2008).
49. Mahasirimongkol, S. *et al.* Genome-wide SNP-based linkage analysis of tuberculosis in Thais. *Genes Immun* **10**, 77–83, <https://doi.org/10.1038/gene.2008.81> (2009).
50. Alcais, A., Abel, L. & Casanova, J. L. Human genetics of infectious diseases: between proof of principle and paradigm. *J Clin Invest* **119**, 2506–2514, <https://doi.org/10.1172/JCI38111> (2009).
51. Tsolaki, A. G. *et al.* Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis. *J Clin Microbiol* **43**, 3185–3191, <https://doi.org/10.1128/JCM.43.7.3185-3191.2005> (2005).
52. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J Clin Microbiol* **35**, 907–914 (1997).
53. Coll, F. *et al.* SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics* **28**, 2991–2993, <https://doi.org/10.1093/bioinformatics/bts544> (2012).
54. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
55. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191, <https://doi.org/10.1093/bioinformatics/btp033> (2009).
56. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731–2739, <https://doi.org/10.1093/molbev/msr121> (2011).
57. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307–321, <https://doi.org/10.1093/sysbio/syq010> (2010).
58. Galtier, N., Gouy, M. & Gautier, C. SEAVIEW and PHYLO\_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12**, 543–548 (1996).
59. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821–829, <https://doi.org/10.1101/gr.074492.107> (2008).
60. Hawkesy, J. *et al.* ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics* **16**, 667, <https://doi.org/10.1186/s12864-015-1860-2> (2015).

## Acknowledgements

We thank all the staffs at the TB/HIV Research Foundation in Chiangrai for their invaluable contribution of provision of information and samples to this project. We are grateful to Prof. Tim Flegel, Zoe Hammatt and Supada Plitphongnaphim for critical comments on the manuscript. The isolates were collected as part of a tuberculosis cohort study supported by various sources, including the Research Program on Emerging and Re-emerging Infectious Diseases currently from the Japan Agency for Medical Research and Development (AMED) and the International Research Funds to RIT by the Ministry of Health and Welfare, Japan. This work has been funded mainly under the Science and Technology Research Partnership for Sustainable Development (SATREPS) project by AMED and the Japan International Cooperation Agency (JICA) and Department of Medical Sciences, Ministry of Public Health, Thailand. It is also supported in part by Sanger Institute, the University of Tokyo, Mahidol University and the National Science and Technology Development Agency, Thailand. PrP is partially supported by the Fogarty International Centre of the National Institutes of Health under award Number D43 TW 009522. TGC is supported by the Medical Research Council UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MC\_PC\_15103). JPh is funded by a BBSRC PhD Studentship. We thank the Wellcome Trust Sanger Institute core and pathogen sequencing and informatics teams who were involved in the sequencing and data assessment. This work was funded in part by the Wellcome Trust (Grant numbers WT096249/Z/11/B, WT088559MA, WT081814/Z/06/Z, and WT098051). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

## Author Contributions

Pr.P., P.A., S.M. drafted the manuscript. J.Pa. oversaw the sequencing. P.A., W.V., N.S., J.Ph., W.R., Pa.P., T.J., S.T. analysed the sequences. A.D., S.N., W.I., P.K., B.C., J.W., H.Y., N.Y. prepared the samples and collected and analysed the patient information. A.D., N.S., T.J. performed the spoligotyping. Pr.P., S.M., T.C., M.H., L.T., K.T., L.T. guided the analysis.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-29986-3>.

**Competing Interests:** The authors declare no competing interests.

**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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018