Whole-genome comparative analysis at the lineage/sublineage level discloses relationships between *Mycobacterium tuberculosis* genotype and clinical phenotype

Andrea Monserrat Negrete-Paz [Equal first author, 1], Gerardo Vázquez-Marrufo [Equal first author, 1], Ma. Soledad Vázquez-Garcidueñas [Corresp. 2]

[1] Centro Multidisciplinario de Estudios en Biotecnología, Facultad de Medicina Veterinaria y Zootecnia, Universidad Michoacana de San Nicolás de Hidalgo, Tarímbaro, Michoacán, Mexico

[2] División de Estudios de Posgrado, Facultad de Ciencias Médicas y Biológicas "Dr. Ignacio Chávez", Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, México

Corresponding Author: Ma. Soledad Vázquez-Garcidueñas
Email address: soledad.vazquez@umich.mx

Background. Human tuberculosis (TB) caused by members of the *Mycobacterium tuberculosis* complex (MTBC) is the main cause of death among infectious diseases worldwide. Pulmonary TB (PTB) is the most common clinical phenotype of the disease, but some patients develop an extrapulmonary (EPTB) phenotype in which any organ or tissue can be affected. MTBC species include nine phylogenetic lineages, with some appearing globally and others being geographically restricted. EPTB can or not have pulmonary involvement, challenging its diagnosis when lungs are not implicated, thus causing an inadequate treatment. Finding evidence of a specific *M. tuberculosis* genetic background associated with EPTB is epidemiologically relevant due to the virulent and multidrug-resistant strains isolated from such cases. Until now, the studies conducted to establish associations between *M. tuberculosis* lineages and PTB/EPTB phenotypes have shown inconsistent results, which are attributed to the strain predominance from specific *M. tuberculosis* lineages/sublineages in the samples analyzed and the use of low-resolution phylogenetic tools that have impaired sublineage discrimination abilities. The present work elucidates the relationships between the MTBC strain lineages/sublineages and the clinical phenotypes of the disease as well as the antibiotic resistance of the strains. **Methods**. To avoid biases, we retrieved the raw genomic reads (RGRs) of all (n=245) the *M. tuberculosis* strains worldwide causing EPTB available in databases and an equally representative sample of the RGRs (n=245) of PTB strains. A multiple alignment was constructed, and a robust maximum likelihood phylogeny based on single-nucleotide polymorphisms was generated, allowing effective strain lineage/sublineage assignment. **Results.** A significant Odds Ratio (OR range: 1.8-8.1) association was found between EPTB and the 1.1.1, 1.2.1, 4.1.2.1 and ancestral Beijing sublineages. Additionally, a significant association between

PTB with 4.3.1, 4.3.3, and 4.5 and Asian African 2 and Europe/Russia B0/W148 modern Beijing sublineages was found. We also observed a significant association of Lineage 3 strains with multidrug resistance (OR 3.8; 95% CI 1.1-13.6), as well as between modern Beijing sublineages and antibiotic resistance (OR 4.3; 3.8-8.6). In this work, it was found that intralineage diversity can drive differences in the immune response that triggers the PTB/EPTB phenotype.

1 Whole-genome comparative analysis at the
2 lineage/sublineage level discloses relationships
3 between *Mycobacterium tuberculosis* genotype and
4 clinical phenotype

6 Andrea Monserrat Negrete-Paz[1], Gerardo Vázquez-Marrufo[1], Ma. Soledad Vázquez-
7 Garcidueñas[2]*

9 [1] Centro Multidisciplinario de Estudios en Biotecnología, Facultad de Medicina Veterinaria y
10 Zootecnia, Universidad Michoacana de San Nicolás de Hidalgo, Tarímbaro, Michoacán, Mexico
11 [2] División de Estudios de Posgrado, Facultad de Ciencias Médicas y Biológicas "Dr. Ignacio
12 Chávez", Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, Mexico

14 *Corresponding Author:
15 Ma. Soledad Vázquez-Garcidueñas[2]
16 Ave. Rafael Carrillo esq. Dr. Salvador González Herrejón S/N, Col. Cuauhtémoc
17 C.P. 58020, Morelia, Michoacán, Mexico
18 E-mail address: soledad.vazquez@umich.mx

21 Abstract

22 **Background.** Human tuberculosis (TB) caused by members of the *Mycobacterium tuberculosis*
23 complex (MTBC) is the main cause of death among infectious diseases worldwide. Pulmonary
24 TB (PTB) is the most common clinical phenotype of the disease, but some patients develop an
25 extrapulmonary (EPTB) phenotype in which any organ or tissue can be affected. MTBC species
26 include nine phylogenetic lineages, with some appearing globally and others being
27 geographically restricted. EPTB can or not have pulmonary involvement, challenging its
28 diagnosis when lungs are not implicated, thus causing an inadequate treatment. Finding evidence
29 of a specific *M. tuberculosis* genetic background associated with EPTB is epidemiologically
30 relevant due to the virulent and multidrug-resistant strains isolated from such cases. Until now,
31 the studies conducted to establish associations between *M. tuberculosis* lineages and PTB/EPTB
32 phenotypes have shown inconsistent results, which are attributed to the strain predominance
33 from specific *M. tuberculosis* lineages/sublineages in the samples analyzed and the use of low-
34 resolution phylogenetic tools that have impaired sublineage discrimination abilities. The present
35 work elucidates the relationships between the MTBC strain lineages/sublineages and the clinical
36 phenotypes of the disease as well as the antibiotic resistance of the strains.
37 **Methods**. To avoid biases, we retrieved the raw genomic reads (RGRs) of all (n=245) the *M.*
38 *tuberculosis* strains worldwide causing EPTB available in databases and an equally

39   representative sample of the RGRs (n=245) of PTB strains. A multiple alignment was
40   constructed, and a robust maximum likelihood phylogeny based on single-nucleotide
41   polymorphisms was generated, allowing effective strain lineage/sublineage assignment.
42   **Results.** A significant Odds Ratio (OR range: 1.8-8.1) association was found between EPTB and
43   the 1.1.1, 1.2.1, 4.1.2.1 and ancestral Beijing sublineages. Additionally, a significant association
44   between PTB with 4.3.1, 4.3.3, and 4.5 and Asian African 2 and Europe/Russia B0/W148
45   modern Beijing sublineages was found. We also observed a significant association of Lineage 3
46   strains with multidrug resistance (OR 3.8; 95% CI 1.1-13.6), as well as between modern Beijing
47   sublineages and antibiotic resistance (OR 4.3; 3.8-8.6). In this work, it was found that
48   intralineage diversity can drive differences in the immune response that triggers the PTB/EPTB
49   phenotype.

50

51

Introduction

53   Tuberculosis (TB) represents the main cause of death among infectious diseases worldwide, with
54   its drug-resistant manifestations constituting a major global health concern (Floyd *et al.*, 2018).
55   Pulmonary TB (PTB) is the most common clinical disease phenotype, but some patients develop
56   an extrapulmonary TB (EPTB) phenotype in which practically any organ or tissue can be
57   affected, including the aggressive manifestations of lymph node and central nervous system TB
58   (Golden & Vikram, 2005). EPTB represents approximately 15% of reported TB cases globally,
59   whereas as many as 40% of TB cases in several high-income countries are EPTB (WHO, 2020).
60   Human TB is caused by members of the *Mycobacterium tuberculosis* complex (MTBC), which
61   have >99% nucleotide sequence identity at the genomic level (Gagneux, 2018). The human-
62   adapted species of the MTBC are *M. tuberculosis sensu stricto* and *Mycobacterium africanum*,
63   which are divided into nine phylogenetic lineages: L1, or Indo-Oceanic; L2, or East Asian; L3,
64   or East African-Indian; L4, or Euro-American; L5, or *M. africanum* West-African 1; L6, or *M.*
65   *africanum* West-African 2; L7, or Ethiopia (Firdessa *et al.,* 2013; Coscolla & Gagneux, 2014);
66   L8, or *M. tuberculosis* from the African Great Lakes *(*Ngabonziza *et al.,* 2020*)*; and the recently
67   described *M. africanum* L9 (Coscolla *et al.,* 2021)*.* Global phylogeographic reconstruction of *M.*
68   *tuberculosis* suggests that each lineage has become specifically adapted to defined human
69   populations (Gagneux, 2018), with some occurring globally and others showing strong
70   geographical restriction. L4, L2, and L3 are the most commonly found lineages; nevertheless, L4
71   is the most widespread lineage worldwide. L3 is mostly found in the Middle East, India, and East
72   Africa, while L2 is found predominantly in East Asia (McHenry *et al.*, 2020). This geographical
73   restriction has been found even at the sublineage level, as in sublineage 4.6/Uganda, found only
74   in Uganda and neighboring countries. This sublineage has been shown to possess highly
75   conserved T cell epitopes and a restricted geographic distribution, suggesting a possible
76   adaptation to a specific human population (Stucki *et al.,* 2016). On the other hand, the high
77   incidence of infections associated with nongeographically restricted strains might imply that
78   these strains are more effective in causing the disease (Malik & Godfrey-Fausset, 2005).

79   Several studies have been conducted to identify possible relationships between *M.*
80 *tuberculosis* phylogenetic lineages and the PTB or EPTB phenotype of the disease (Feng *et al*.,
81 2008; Click *et al*., 2012), but the results show a lack of consistency. Different factors contribute
82 to explaining the observed discrepancies among the conducted studies in an attempt to establish
83 genotype-phenotype relationships. In the first instance, there were differences in the sample size
84 (Coscolla & Gagneux, 2014), as well as in the nonhomogeneous distribution of lineages, among
85 the set (Kato-Maeda & Nahid, 2012) of analyzed strains. Moreover, biased associations might
86 arise due to a lack of data or failure to control for possible confounders associated with known
87 risk factors for EPTB, such as human immunodeficiency virus (HIV) infection comorbidity in
88 patients from whom strains were isolated (Coscolla & Gagneux, 2014). Furthermore, studies use
89 different operational definitions for EPTB (Kato-Maeda & Nahid, 2012)*,* and some lack
90 appropriate tools to index genomic diversity and classify strains into lineages in some studies
91 (Coscolla & Gagneux, 2010). Whole-genome comparative analysis has allowed the
92 categorization of *M. tuberculosis* lineages into sublineages using single-nucleotide
93 polymorphism (SNP) analysis (Coll *et al*., 2014; Stucki *et al*., 2016). This subtle level of strain
94 differentiation suggests that some sublineages might drive the observed associations of an *M.*
95 *tuberculosis* genotype with a specific disease phenotype, such as EPTB (Feng *et al*., 2008), but
96 an analysis to provide evidence in support of this hypothesis has not been conducted.
97 Interestingly, the frequency of the sublineages assigned to isolated strains from the East Asian
98 (L2) lineage differs among populations settled in different geographical areas, which might also
99 explain why some studies associate L2 lineage strains with EPTB (Feng *et al*., 2008), whereas
100 others associate it with PTB (Dale *et al*., 2005), and still others do not find any association at all
101 (Svensson *et al.,* 2011). Thus, intralineage diversity requires a detailed exploration to clarify
102 disease phenotype variation and its relationship with MTBC genotypes (Kato-Maeda & Nahid,
103 2012). Additionally, genomic evidence has revealed a close relationship among specific *M.*
104 *tuberculosis* lineages and sublineages with drug resistance (Wang *et al*., 2015), a relevant public
105 health phenotype that might hinder successful TB treatment. Horizontal transfer of drug
106 resistance genes has not been reported for *M. tuberculosis*, but resistance mostly arises from
107 chromosomal mutations under the selective pressure of antibiotic use (Nguyen, 2016). Thus, the
108 characterization of mutations associated with resistance phenotypes in strains with different
109 genotypes can help to reveal lineage-/sublineage-specific microevolutionary processes of
110 epidemiological relevance. Drug-resistance-associated mutations have been hypothesized to
111 modify strain fitness and the ability of a strain to cross the blood-brain barrier, causing EPTB,
112 specifically tuberculous meningitis (Faksri *et al*., 2018).
113   Unclarified relationships of lineages/sublineages with *M. tuberculosis* PTB/EPTB disease
114 and drug resistance phenotypes hinder the generation of adequate epidemiological transmission
115 chains and timely successful treatments. Therefore, this work aims to elucidate relationships
116 between strain genotypes at the lineage/sublineage level and clinical disease phenotypes as well
117 as antibiotic resistance. We retrieved the raw read datasets from the genomes of all the *M.*
118 *tuberculosis* strains causing EPTB worldwide and deposited in databases. The raw datasets of the

119   same number of genomes of strains causing PTB isolated from the same countries of origin as
120   EPTB strains were selected to avoid sources of possible biases related to previous studies
121   originating from (i) a low number of analyzed strains, (ii) unequal PTB/EPTB strains analyzed,
122   or (iii) differences in regional clinical TB phenotype incidences. The epidemiological and public
123   health relevance of the identified relationships is discussed.
124
125

126   **Materials & Methods**
127

128   **Data retrieval**
129   A total of 490 raw datasets of genome sequence reads were retrieved, which corresponded to 245
130   *M. tuberculosis* strains causing PTB and 245 strains causing EPTB (Supplemental Table 1). We
131   retrieved all available genomes of EPTB strains deposited in the NCBI-SRA database. Such
132   genomes correspond to clinical cases for which the major site of infection reported in databases
133   was not pulmonary or miliary and an additional infection site either was not specified or was
134   specified but not as pulmonary or miliary (Click *et al.*, 2012). The first criterion for the genome
135   selection of the PTB strains was the country of the isolation of the EPTB strains for which the
136   genomes were available. The objective of such a criterion was to analyze the genomes of *M.
137   tuberculosis* phenotypes from similar clinical and human population backgrounds. The second
138   criterion to select PTB strains was a negative HIV status reported in the associated metadata. In
139   fact, this second criterion was applied for both PTB and EPTB genomes. Additionally, due to the
140   large number of available genomes of PTB strains from different countries, it was possible to
141   discard other comorbidities. All these criteria were used to avoid possible biases generated by
142   population-genotype associations and HIV, or other comorbidities.
143   The sequence quality of the FASTQ reads was checked using FASTQC
144   (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and subsequently filtered to a Phred
145   score of 30 using TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).
146

147   *In silico* **typing**
148   *In silico* spoligotyping was performed using SpoTyping program version 2.0 (Xia, Teo & Ong,
149   2016) for next-generation sequencing reads with default parameters. To determine lineage, the
150   TB INSIGHT (http://tbinsight.cs.rpi.edu) server was used based on the identified spoligotypes.
151

152   **Phylogenetic reconstruction**
153   To find SNPs for phylogenetic reconstruction (Homolka *et al.*, 2012; Coll *et al.*, 2014; Merker *et
154   al.*, 2015), the sequencing reads of the studied strains were aligned to the reference strain of *M.
155   tuberculosis* H37Rv (accession no. NC_000962.3) using the MTBseq program version 1.0.3
156   (Kohl *et al.*, 2018) with default values. A frequency of allelic variation greater or equal to 75%
157   and with a phred value > 20 was used. Strains where the percentage of reads mapped against the
158   reference genome was less than 80 were excluded due to possible contamination. Also, strains

159   with a median depth coverage > 30x were removed. The minimum coverage depth to support a
160   SNP was 8x. For lineage and sublineage assignment, variant positions belonging to repeated
161   regions and resistance genes were excluded. Of the remaining variant positions, those where data
162   quality is below thresholds in > 5% of samples were discarded (Jajou *et al*, 2019). The
163   sublineage phylogeny was rooted using a *Mycobacterium microti* strain (SRR2667442).
164   Phylogenetic inferences were conducted with the maximum likelihood (ML) criterion using the
165   IQ-TREE package (Nguyen *et al*., 2015) with a general time-reversible (GTR) model of
166   nucleotide substitution and a gamma model of rate heterogeneity. Phylogenetic trees were
167   constructed based on 1,000 bootstrap replicates, and their visualization was performed using
168   iTOL (Letunic & Bork, 2016).
169
170   ***In silico* determination of drug resistance**
171   Raw FASTQ sequencing files were uploaded to TB-Profiler version 3.0.4 (Coll *et al*., 2015), a
172   tool to determine *in silico* drug resistance. TB-Profiler can determine genotypic drug resistance
173   by aligning raw sequences against the reference genome *M. tuberculosis* H37Rv in order to
174   identify mutations (1541 SNPs and indels) previously associated with phenotypic drug-resistance
175   from a curated database. This package also determines the *M. tuberculosis* lineage based on a 90-
176   SNP barcode. The TB-Profiler-predicted resistance mutations were validated using the results of
177   MTBseq, which reports a list of mutations in genes associated with antimicrobial resistance for
178   every processed strain.
179
180   **Data analysis**
181   Data entry and statistical analyses were performed in SPSS version 16 (SPSS Inc., Illinois,
182   USA). Univariate analyses were performed using two-tailed Fisher's exact test to estimate the
183   association between each variable (*M. tuberculosis* lineage, sublineage or genotypic resistance)
184   and extrapulmonary tuberculosis relative to pulmonary tuberculosis. Additionally, each variable
185   was compared to every anatomical site of TB disease (central nervous system, bones and joints,
186   lymph nodes, and the genitourinary system). Odds ratios with 95% confidence intervals were
187   considered an effect size of the association. To confirm that the association was not an artifact of
188   demographic differences between the geographic regions, a logistic regression model was
189   performed in which the EPTB group was included as a dependent variable. The variables
190   geographic region (the country of isolation), sublineage and genotypic resistance were included
191   in the model, and adjusted ORs with 95% confidence intervals were calculated.
192
193
194   # Results
195   **Anatomical site of infection of strains selected for comparative genomics analysis**
196   A total of 490 raw genome datasets of *M. tuberculosis* strains isolated from seven different
197   countries from individuals with a negative HIV infection status (Supplemental Table 1) were
198   genotyped. The selected strains included those from countries where specific lineages

199   predominate according to the SITVIT website (http://www.pasteur-
200   guadeloupe.fr:8081/SITVIT_ONLINE/), such as Thailand and Indonesia for the East Asian
201   lineage and India for the East African-Indian lineage. Ninety percent of the selected strains were
202   isolated from Indonesia, Thailand, and Russia. This high percentage is since most of the EPTB
203   strains deposited in the NCBI-SRA database were isolated from these countries. Unfortunately,
204   EPTB strains from African and American countries were not included in the conducted analysis
205   because of the lack of genomes in the NCBI-SRA database until the completion of this study.
206   The EPTB phenotype strains were clustered into five major groups according to the
207   anatomical site of infection, with a predominance of the central nervous system (72.24%). Other
208   anatomical sites of the disease were bones and joints (17.14%), lymph nodes (4.49%), and the
209   genitourinary system (2.86%); the remaining 3.27% were from sites specified as having only
210   EPTB strains in databases.
211
212   **Distribution and association of lineages and genotypes with clinical phenotype**
213   The mean coverage obtained for the analyzed strain set was 109.82, whereas the mean
214   percentage of mapped reads was 98.28, resulting in good genome coverage. The strains were
215   assigned to a major *M. tuberculosis* genetic lineage according to the spoligotype using the TB
216   lineage search option of the TB insight web server. These results correlated with the clustering
217   pattern generated by the SNP-based phylogeny. The phylogenetic analysis of *M. tuberculosis*
218   strains from PTB and EPTB phenotypes is shown in Figure 1. The predominant lineage of the
219   analyzed strains was East Asian (54.28%), followed by Euro-American (34.70%), Indo-Oceanic
220   (9.6%), and East African-Indian (1.42%). The EPTB case percentage differed among these
221   phylogenetic lineages, with 15.51% for Indo-Oceanic, 48.98% for East Asian, 2.45% for East
222   African-Indian, and 33.06% for Euro-American. The same variation in phylogenetic lineages
223   was observed for strains from PTB cases, with 59.59%, 36.32%, 3.67%, and 0.40% for the East
224   Asian, Euro-American, Indo-Oceanic, and East African-Indian lineages, respectively. Fisher's
225   exact test showed that PTB was significantly associated with East Asian lineage strains (OR 1.5;
226   95% CI 1.1–2.2) and that EPTB was significantly associated with the Indo-Oceanic lineage
227   strains (OR 4.8; 95% CI 2.2–10.1) (Supplemental Table 2). After logistic regression analysis, the
228   associations between L1 and EPTB (OR 4.4; 95% CI 1.9-10.0 P=0.000) and L2 and PTB (OR
229   1.4; 95% CI 1.1-1.9) strains was confirmed, thus eliminating any potential bias due to the
230   isolation regions of *M. tuberculosis* strains (Supplemental Table 2).
231   We searched for a relationship between the phylogenetic lineages found in the analyzed
232   strains and anatomical sites of infection, including lungs, lymph nodes, the genitourinary system,
233   the central nervous system, and bones and joints. The results showed that strains belonging to the
234   Indo-Oceanic lineage were significantly associated with central nervous system infection (OR
235   3.9; CI 95% 2.1-7.4) (Supplemental Table 3), whereas those of the East Asian lineage were
236   significantly associated with bone and joint infection (OR 1.5; 95% CI 1.3-1.9) and with PTB
237   strains, as previously described. Further division of identified lineages was conducted using SNP
238   analysis, distinguishing 27 unique sublineages, with the predominant sublineages being 2.2.1

239   (n=218 strains), 2.2.1.1 (n=25), 4.1.2.1 (n=18), 4.3.1 (n=25), 4.8 (n=51), and 1.2.1 (n=22). As
240   expected, statistical analysis showed a significant relationship of some of these sublineages with
241   major infection sites. In this way, sublineage 1.1.1 was associated with central nervous system
242   infection (OR 2.8; 95% CI 1.0–7.7), and the same relationship was observed for sublineages
243   1.2.1 (OR 6.7; 95% CI 2.0-22.1) and 4.1.2.1 (OR 6.3; 95% CI 1.9-21.1). On the other hand,
244   sublineages 4.3.1 (OR 2.9; 95% CI 1.2-7.7), 4.3.3 (OR 1.8; 95% CI 1.4-2.2), and 4.5 (OR 3.8;
245   95% CI 1.0-13.8) were associated with the pulmonary phenotype.
246   A deeper whole-genome SNP-based classification to discriminate the proto-Beijing from
247   the Beijing strains allowed us to identify the outbreak sublineages of Asia Ancestral 1, Asia
248   Ancestral 3, Asian African 2, Asian African 2/RD142, Asian African 3, Pacific RD150, Central
249   Asia, and Europe/Russia B0/W148 and a group of modern unclassified strains using informative
250   genetic markers found in the genomes of the studied strains (Figure 2). Of all these genotypes,
251   the Asian African 2 sublineage (OR 2.3; 95% CI 1.1-5.2) and Europe/Russia B0/W148 outbreak
252   sublineage (OR 2.7; 95% CI 1.3-5.4) were significantly associated with PTB. In the same way,
253   the Ancestral Beijing sublineages (Asia Ancestral 1 and Asia Ancestral 3) (OR 2.4; 95% CI 1.2-
254   4.8) and Central Asia subgroup (OR 8.2; 95% CI 2.9-22.9) were significantly associated with
255   EPTB.
256   The most frequent spoligotype among L2 strains (85.1%) was SIT1 (Supplemental Table
257   4). SIT 53 and SIT 19 were the most frequent spoligotypes for the L4 (22.9%) and L1 (27.6%)
258   strains, respectively. For the L3 strains, 42% had an unclassified SIT spoligotype
259   (703775740003771). Spoligotyping allowed the identification of twelve major genotypic
260   families, including Beijing (53.7%), Central Asia (CAS) (1.4%), EAI (9.8%), Haarlem (4.4%),
261   Latin American–Mediterranean (LAM) (7.5%), S (0.2%), T (19.8%), X (0.4%) and Family33-36
262   (2%). Interestingly, 13.26% of the strains corresponded to spoligotypes not previously reported
263   in the SITVIT database (Supplemental Table 4). A significant association was found between the
264   EAI2 family (OR 6.7; 95% CI 2.0-22.1) and EPTB, of which 13 strains caused tuberculous
265   meningitis and two of them led to lymph node infection.
266
267   **Mutations associated with drug-resistant TB**
268   A total of 180 previously reported mutations distributed in 18 genes known to confer resistance
269   to first- and second-line drugs for TB treatment were identified (Supplemental Table 5).
270   Isoniazid was the antibiotic with the highest predicted resistance in 38.77% of the studied strains,
271   followed by rifampicin in 33.46%, streptomycin in 31.02% and ethambutol in 27.14%.
272   The most common mutation associated with isoniazid resistance was *katG* Ser315Thr, which
273   was found in 82.63% of the genotypic resistant strains. *rpoB* Ser450Leu was found in 52.43% of
274   the rifampicin-resistant strains, *rpsL* Lys43Arg in 72.36% of streptomycin-resistant strains and
275   *embB* Met306Val in 33.83% of ethambutol-resistant strains.
276   Extrapulmonary strains showed greater diversity of mutations but appeared less
277   frequently than pulmonary strains. Mutations *gyrA* (Ala90Val) and *thyX* (Glu40Gly) were found
278   only in pulmonary strains. On the other hand, mutations in the *rpoB* (Leu430Pro), *rrs* (907

279   A>C), *rpsL* (Lys88Thr), *embA* (16C>G), *embB* (Ala356Val, Ser347Ile), *ethA* (Gln269*,

280   Gly385Asp), *katG* (589 insGG), and *pncA* (Ala3Glu, Cys138Arg, Cys72Arg, Gln10Pro,

281   His51Asp, Gly97Asp, Pro54Leu, Thr135Pro) genes were found only in extrapulmonary strains.

282   There were no statistically significant differences among these variations with the tuberculosis

283   clinical phenotype.

284   Differences between the distribution of drug-resistant (26.9% PTB, 15.9% EPTB) and

285   drug-sensitive (23.1% PTB, 34.1% EPTB) strains were observed in the EPTB and PTB groups.

286   Four resistance profiles were determined: sensitive, drug resistance (DR), multidrug resistance

287   MDR, and extensive drug resistance (XDR). The frequency for each EPTB/PTB group is shown

288   in Supplemental Table 2. The bone and joint group presented the highest number of strains with

289   antibiotic resistance: 76.1% were resistant to isoniazid, 73.8% to streptomycin, 66.6% to

290   rifampicin, 54.7% to ethambutol, 42.8% to pyrazinamide and ethionamide, 14.2% to

291   fluoroquinolones and 9.5% to aminoglycosides and paraaminosalicylic acid (Supplemental Table

292   6). A total of 13.5% of central nervous system strains showed resistance to isoniazid, and 54.4%

293   of lymph node strains were resistant to rifampicin. Sensitive strains were statistically associated

294   with EPTB (OR: 1.7; 95% CI 1.3-2.3), and extensive drug resistance (XDR; OR: 6.0; 95% IC

295   2.5-14.4) was more strongly associated with PTB strains than with EPTB strains. After logistic

296   regression analysis including the geographic region of isolation, XDR strains were still

297   associated with pulmonary rather than extrapulmonary disease (OR: 6.0; 95% IC 2.3-14.4

298   P=0.000). We also observed a significant association of Lineage 3 strains with MDR (OR: 3.8;

299   95% CI 1.1-13.6) and between the L2 lineage and XDR resistance (OR: 8.1; 5.4-8.6).

300

301

Discussion

303   In this work, we used both raw genome datasets and whole-genome SNP-based phylogeny to

304   genotype the same number of PTB and EPTB *M. tuberculosis* strains. All the strains were

305   included in one of the six lineages defined by Coll (Coll *et al*., 2014). When seeking lineage-

306   clinical phenotype relationships, we found that the East Asian 'modern' *M. tuberculosis* lineages

307   were associated with PTB, whereas the Indo-Oceanic so-called an 'ancient lineage' were

308   associated with EPTB. It is known that strains from 'modern' lineages induced a slighter

309   inflammatory response than those from 'ancient' lineages, which has been related to a selective

310   advantage of strains from 'modern' lineages, resulting in impaired bacterial control by the host,

311   faster disease progression, and enhanced transmission (Portevin *et al*., 2011). These clinical

312   differences contribute to explaining the association of the East Asian lineage with the PTB

313   phenotype, supporting the hypothesis that since this is the disease contagious phenotype, patients

314   infected with strains of such a lineage are prone to develop pulmonary disease at a higher

315   frequency than patients infected with strains from other genotypes, which is consistent with the

316   increased transmissibility of strains from this lineage (Thwaites *et al*., 2018).

317   It has been proposed that strains belonging to the Indo-Oceanic lineage are 'less virulent'

318   than those from other lineages and cause a specific exacerbated inflammatory response

319 (Chakraborty *et al*., 2018), which might be attributed to the presence of unique cell envelope
320 lipids in this lineage, such as phenol phthiocerol dimycocerosate (Krishnan *et al*., 2011).
321 Nevertheless, it is still unknown whether these differences in mycobacterial cell envelope lipid
322 composition can explain lineage-related phenotypic differences such as the EPTB phenotype.
323 Interestingly, the percentage of EPTB cases associated with the East African-Indian lineage was
324 as high as 85%, but this finding was not statistically significant due to the small number of EPTB
325 raw genomes (n=7) from this lineage available in databases. In this regard, the present work
326 reveals database gaps relevant for its public health and epidemiological implications, i.e., the
327 need to include more East African-Indian lineage genomes to clarify its relationships with EPTB.
328   In the search for a detailed relationship between phylogenetic lineages and the anatomical
329 site of infection, strains belonging to the East Asian lineage were associated with the infection of
330 both bones and joints and PTB. The East Asian lineage comprises two major clades or
331 sublineages, designated proto-Beijing (2.1) and Beijing (2.2) (Ajawatanawong *et al*., 2019).
332 Sublineage 2.2, or the Beijing family, as defined by spoligotyping, is composed of several
333 sublineages broadly categorized into the ancestral and modern Beijing strains (Mokrousov *et al*.,
334 2005). A subtle SNP-based classification was recently proposed that allows the discrimination of
335 proto-Beijing from Beijing strains (Shitikov *et al*., 2017). Such classification divides the Beijing
336 group into the ancestral Beijing clade and the modern Beijing clade, which comprises two
337 groups, one including three strains (Asia Ancestral 1, Asia Ancestral 2, Asia Ancestral 3) and the
338 other including seven strains (Asian African 1, Asian African 2, Asian African 2/RD142, Asian
339 African 3, Pacific RD150, Europe/Russia B0/W148 outbreak and Central Asia). This
340 classification allows us to associate the Asian African 2 sublineage and Europe/Russia B0/W148
341 outbreak sublineage with PTB, as well as the ancestral Beijing sublineages and Central Asia
342 subgroup with EPTB. Differences in the pathogenicity of Beijing sublineages have been
343 previously reported (Feng *et al.,* 2008), and sublineage specific patterns of induced cytokine
344 production by macrophages have also been observed (Sarkar *et al*., 2012). In this regard, a
345 macrophage infection model revealed that ancestral Beijing strains induce a higher production of
346 the proinflammatory cytokines TNF-α and IL-6 than the modern Beijing sublineage (Chen *et al*.,
347 2014). High IFN-γ expression and cytokine production have also been reported in peripheral
348 blood mononuclear cells of ancestral Beijing strains (Faksri *et al*., 2014). Such results suggest
349 that the ancestral Beijing strains are as highly immunogenic (Kato-Maeda *et al*., 2012) as
350 Lineage 1 strains (EAI) (Rakotosamimanana *et al*., 2010). Interestingly, these two sublineages
351 (EAI and ancestral Beijing) were associated with the EPTB phenotype in this study. In contrast,
352 the other sublineage associated with EPTB in this work, 4.1.2.1, has been reported to induce
353 cytokine levels similar to those of H37Rv (Haarlem family strains), which are associated with a
354 low immune response (Wang *et al*., 2010). This supports the idea that strains of different
355 sublineages vary by many phenotypes, such as the tendency to develop drug resistance, virulence
356 levels, and immune response, which influence disease severity and clinical presentation.
357   Interestingly, sublineages 4.1.2.1 and ancestral Beijing, belonging to Lineages 4 and 2,
358 respectively, show a high prevalence worldwide, representing more than 50% of the strains in

359   certain areas and/or subpopulations (Ajawatanawong *et al.*, 2019). In contrast, Lineage 1 (EAI)
360   strains are commonly reported in countries around the Indian Ocean and are one of the most
361   geographically restricted families. However, EAI strains have been reported in lower percentages
362   in countries such as the Netherlands, Australia, the USA, Sweden, Saudi Arabia, Tunisia,
363   Taiwan, Panama, and Mexico. The East Asian India 2 spoligotype of Lineage 1 corresponds to
364   the Nonthaburi (EAI2-Nonthaburi) genotype and the Manila (EAI2-Manila) genotype (Couvin,
365   Reynaud & Rastogi, 2019). Recently, Coker (Coker *et al.* 2016) reported a polymorphism in the
366   genome of Nonthaburi strains from three patients with tuberculous meningitis, consisting of a
367   500 bp deletion covering ppe50 that was not present in the reference strain *M. tuberculosis*
368   H37Rv. They reported three mutations (T28910C, C1180580T and, C152178T) until now found
369   only in these meningeal Nonthaburi strains. These mutations could represent part of the genetic
370   background that could be shared by strains that cause EPTB and that also belong to the EAI,
371   4.1.2.1, and ancestral Beijing sublineages. Nonetheless, further investigation is required to
372   determine whether these mutations are shared, which could be the functional consequences and
373   the probable relation of such polymorphisms with TB disease phenotype.
374   Regarding the genotype relationship with antibiotic resistance, to the best of our
375   knowledge, a significant association of Lineage 3 with drug resistance found here has not been
376   previously reported. This result might be useful to optimize TB treatment in geographical areas
377   where this lineage is frequent. Studies from Asia, Europe and Africa have shown varying
378   associations between drug resistance and MTB lineages (Coscolla & Gagneux, 2010; Singh *et*
379   *al.*, 2015); however, as was found here, Beijing strains have been associated with MDR and
380   XDR in several cases (Rodríguez-Castillo *et al.*, 2017). However, it must be highlighted that the
381   lack of experimental drug resistance assays is a limitation of this study. The bioinformatic
382   approach to determine the drug resistance of the studied isolates is based on software packages
383   that use different sets of mutations in their analysis. This generates a variation in the drug-
384   resistant genotypes obtained with the different bioinformatic tools with the consequent risk of
385   over- or underestimating the true drug-resistant behavior of the analyzed strains. The statistical
386   inferences that depend on this determination are also subject to such biases as the association of
387   drug resistance with affected organs or mycobacterial lineages. Thus, it is strongly recommended
388   to perform the experimental determination of the drug resistance phenotype of the strains in
389   which the genome will be deposited in public databases.
390   The advantages of the present study regarding previous works seeking relationships of
391   the *M. tuberculosis* genotype with clinical phenotype include the use of a significantly higher
392   number of strains, an equal number of EPTB and PTB strains, a clear assignment of the strains
393   considered to be associated with EPTB, and the exclusion of the strains from HIV patients, with
394   the latter being a controlled variable, a caution not commonly considered in similar works.
395   However, other possible confounders or known risk factors for EPTB could not be considered in
396   the analysis conducted here, partly due to the lack of additional metadata of clinical information
397   in the database where the genomes of the strains were retrieved. This lack of metadata is not a
398   source of bias in the relationships found here because it has been shown that there is an

399 independent association between lineages and EPTB rather than ethnic factors after a stratified
400 analysis (Click *et al.,* 2012). Additionally, it must be noted that the inclusion of genomes
401 previously published that can be generated for different purposes can generate a bias by sampling
402 protocols to seek specific pathogen genotypes or tuberculosis phenotypes. This was a not
403 controlled variable of this study.
404   These results support the assertion that the relationship between sublineages and clinical
405 disease phenotypes is not attributable to regional-specific factors; therefore, if this association
406 exists, it does not depend only on unknown clinical factors. However, the contribution of the
407 genetic background of the host to the occurrence of the different clinical phenotypes of
408 tuberculosis should not be ruled out. Several polymorphisms associated with EPTB have been
409 previously reported in specific human populations (Fox *et al.*, 2014). Therefore, associations
410 found here must be further tested with additional epidemiological data to clarify the possible
411 relationship between the intrinsic pathogen sublineage characteristics and host factors as genetic
412 background for the establishment of the infection in a particular organ.
413
414

415 **Conclusions**
416 Overall, the obtained results suggest that intralineage diversity could drive differences in the
417 immune response that trigger the different clinical phenotypes of tuberculosis. The immune
418 response caused by ancient lineage (L1) and ancestral Beijing strains could be similar, eliciting a
419 high inflammatory response. Our results highlight the need both to analyze the genomic
420 background shared by these strains and to perform *in vitro*/*in vivo* studies that help to elucidate
421 the mechanism by which they could disseminate through the body, causing extrapulmonary
422 disease. We demonstrated that the lack of consistency regarding clinical phenotype-strain
423 genotype associations in the results obtained by previous studies is partially due to the use of
424 inadequate phylogenetic classification tools, which do not allow discrimination between
425 sublineages. Additionally, the present results were not biased by the predominance of a specific
426 lineage/sublineage in a specific human population because we included all the EPTB strains
427 available in databases and an evenly representative sample of PTB strains. Biases originating
428 from unknown sampling procedures of the strains for which genomes are available in databases
429 and due to unknown phenotypic drug resistance are the main limitations of the present study.
430
431

432 **Acknowledgements**
433 We thank CONACYT-Mexico for scholarship No. 701889 granted to A.M.N.P., and
434 Coordinación de la Investigación Científica UMSNH research program; We appreciate the
435 criticism of Dr. Sebastien Gagneux and an anonymous reviewer that contributed to significantly
436 improve the manuscript.
437

438 **References**

439   Ajawatanawong P, Yanai H, Smittipat N, Disratthakit A, Yamada N, Miyahara R, Nedsuwan S,
440   Imasanguan W, Kantipong P, Chaiyasirinroje B, Wongyai J, Plitphonganphim S,
441   Tantivitayakul P, Phelan J, Parkhill J, Clark TG, Hibberd ML, Ruangchai W,
442   Palittapongarnpim P, Juthayothin T, Thawornwattana Y, Viratyosin W, Tongsima S,
443   Mahasirimongkol S, Tokunaga K, Palittapongarnpim P. 2019. A novel Ancestral Beijing
444   sublineage of *Mycobacterium tuberculosis* suggests the transition site to Modern Beijing
445   sublineages. *Scientific Reports* 9: 1-12 DOI:10.1038/s41598-019-50078-3
446   Chakraborty P, Kulkarni S, Rajan R, Sainis K. 2018. *Mycobacterium tuberculosis* strains from
447   ancient and modern lineages induce distinct patterns of immune responses. *Journal of*
448   *Infection in Developing Countries* 11: 904–911 DOI:10.3855/jidc.8596
449   Chen YY, Chang JR, Huang WF, Hsu SC, Kuo SC, Sun JR, Dou HY. 2014. The pattern of
450   cytokine production in vitro induced by ancient and modern Beijing *Mycobacterium*
451   *tuberculosis* strains. *PLoS ONE* 9: 1-7 DOI:10.1371/journal.pone.0094296
452   Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. 2012. Relationship between
453   *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis.
454   *Clinical Infectious Disease* 54: 211–219 DOI:10.1093/cid/cir788
455   Coker OO, Chaiprasert A, Ngamphiw C, Tongsima S, Regmi SM, Clark TG, Ong RT, Teo YY,
456   Prammananan T, Palittapongarnpim P. 2016. Genetic signatures of *Mycobacterium*
457   *tuberculosis* Nonthaburi genotype revealed by whole genome analysis of strains from
458   tuberculous meningitis patients in Thailand. *PeerJ* 4: 1-20 DOI:10.7717/peerj.1905
459   Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain
460   A, Martin N, Clark TG. 2014. A robust SNP barcode for typing *Mycobacterium*
461   *tuberculosis* complex strains. *Nature communications* 5: 1-5 DOI:10.1038/ncomms5812
462   Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, Mallard
463   K, Nair M, Miranda A, Alves A, Perdigão J, Viveiros M, Portugal I, Hasan Z, Hasan R,
464   Glynn JR, Martin N, Pain A, Clark TG. 2015. Rapid determination of anti-tuberculosis
465   drug resistance from whole-genome sequences. *Genome Medicine* 7: 51–61
466   DOI:10.1186/s13073-015-0164-0
467   Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, Borrell S, Otchere ID,
468   Asante-Poku A, Asare P, Sánchez-Busó L, Gehre F, Sanoussi CN, Antonio M, Affolabi
469   D, Fyfe J, Beckert P, Niemann S, Alabi AS, Grobusch MP, Kobbe R, Parkhill J, Beisel C,
470   Fenner L, Böttger EC, Meehan CJ, Harris SR, de Jong BC, Yeboah-Manu D, Brites D.
471   2021. Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex
472   evolutionary history. *Microbial Genomics* 7:1-14 DOI:10.1099/mgen.0.000477
473   Coscolla M, Gagneux S. 2010. Does *M. tuberculosis genomic* diversity explain disease
474   diversity?. *Drug Discovery Today Disease Mechanisms* 7: 1–26
475   DOI:10.1016/j.ddmec.2010.09.004
476   Coscolla M, Gagneux S. 2014. Consequences of genomic diversity in *Mycobacterium*
477   *tuberculosis*. *Seminars in Immunology* 26: 431–44 DOI:10.1016/j.smim.2014.09.012
478   Couvin D, Reynaud Y, Rastogi N. 2019. Two tales: worldwide distribution of Central Asian
479   (CAS) versus ancestral East-African Indian (EAI) lineages of *Mycobacterium*
480   *tuberculosis* underlines a remarkable cleavage for phylogeographical, epidemiological
481   and demographical characteristics. *PLoS ONE* 14: 1-20
482   doi.org/10.1371/journal.pone.0219706

483 Dale JW, Bothamley GH, Drobniewski F, Gillespie SH, McHugh TD, Pitman R. 2005. Origins
484   and properties of *Mycobacterium tuberculosis* strains in London. *Journal of Medical*
485   *Microbiology* 54: 575–582 DOI:10.1099/jmm.0.45959-0
486 Faksri K, Chaiprasert A, Pardieu C, Casali N, Palaga T, Prammananan T, Palittapongarnpim P,
487   Prayoonwiwat N, Drobniewski F. 2014. Heterogeneity of phenotypic characteristics of
488   the modern and ancestral Beijing strains of *Mycobacterium tuberculosis*. *Asian Pacific*
489   *Journal of Allergy and Immunology* 32: 124–132 DOI:10.12932/AP0361.32.2.2013
490 Faksri K, Xia E, Ong RT, Tan JH, Nonghanphithak D, Makhao N, Thamnongdee N,
491   Thanormchat A, Phurattanakornkul A, Rattanarangsee S, Ratanajaraya C, Suriyaphol P,
492   Prammananan T, Teo YY, Chaiprasert A**.** 2018. Comparative whole-genome sequence
493   analysis of *Mycobacterium tuberculosis* isolated from tuberculous meningitis and
494   pulmonary tuberculosis patients. *Scientific Reports* 8: 1–10 DOI:10.1038/s41598-018-
495   23337-y
496 Feng JY, Su WJ, Tsai CC, Chang SC. 2008. Clinical impact of *Mycobacterium tuberculosis* W-
497   Beijing genotype strain infection on aged patients in Taiwan. *Journal of Clinical*
498   *Microbiology* 46: 3127–3129 DOI:10.1128/JCM.01132-08
499 Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, Gadisa E, Kiros T, Habtamu M,
500   Hussein J, Zinsstag J, Robertson BD, Ameni G, Lohan AJ, Loftus B, Comas I, Gagneux
501   S, Tschopp R, Yamuah L, Hewinson G, Gordon SV, Young DB, Aseffa A. 2013.
502   Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia.
503   *Emerging Infectious Disease* 19: 460-463 DOI:10.3201/eid1903.120256.
504 Floyd K, Glaziou P, Zumla A, Raviglione M. 2018. The global tuberculosis epidemic progress in
505   patient care, prevention and control efforts in year 3 of the End TB Era. *Lancet*
506   *Respiratory Medicine* 6: 299–314 DOI:10.1016/S2213-2600(18)30057-2
507 Fox GJ, Sy DN, Nhung NV, Yu B, Ellis MK, Van Hung N, Cuong NK, Thi Lien L, Marks GB,
508   Saunders BM, Britton WJ. 2014. Polymorphisms of SP110 are associated with both
509   pulmonary and extra-pulmonary tuberculosis among the Vietnamese. *PLoS One*. 9: 1-9
510   DOI:10.1371/journal.pone.0099496.
511 Gagneux S. 2018. Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews*
512   *Mycrobiology* 16: 202–213 DOI: 10.1038/nrmicro.2018.8
513 Golden MP, Vikram HR. 2005. Extrapulmonary tuberculosis: an overview. *American Family*
514   *Physician* 72: 1761–1768.
515 Homolka S, Projahn M, Feuerriegel S, Ubben T, Diel R, Nübel U, Niemann S. 2012. High
516   resolution discrimination of clinical *Mycobacterium tuberculosis* complex strains based
517   on single nucleotide polymorphisms. *PLoS ONE* 7:1-11
518   DOI:10.1371/journal.pone.0039855
519 Jajou R, Kohl TA, Walker T, Norman A, Cirillo DM, Tagliani E, Niemann S, de Neeling A,
520   Lillebaek T, Anthony RM, van Soolingen D. 2019. Towards standardisation: comparison
521   of five whole genome sequencing (WGS) analysis pipelines for detection of
522   epidemiologically linked tuberculosis cases. *Eurosurveillance*. 24: 1-10 DOI:
523   10.2807/1560-7917.ES.2019.24.50.1900130.
524 Kato-Maeda M, Nahid P. 2012. *Mycobacterium tuberculosis* lineage-what's in your lungs?.
525   *Clinical Infectious Disease* 54: 220–224 DOI:10.1093/cid/cir795
526 Kato-Maeda M, Shanley CA, Ackart D, Jarlsberg LG, Shang S, Obregon-Henao A, Harton M,
527   Basaraba RJ, Henao-Tamayo M, Barrozo JC, Rose J, Kawamura LM, Coscolla M,
528   Fofanov VY, Koshinsky H, Gagneux S, Hopewell PC, Ordway DJ, Orme IM. 2012.

529   Beijing sublineages of *Mycobacterium tuberculosis* differ in pathogenicity in the guinea
530   pig. *Clinical Vaccine Immunology* 19: 1227–1237 DOI:10.1128/CVI.00250-12
531 Kohl TA, Utpatel C, Schleusener V, De Filippo MR, Beckert P, Cirillo DM, Niemann S. 2018.
532   MTBseq: A comprehensive pipeline for whole genome sequence analysis of
533   *Mycobacterium tuberculosis* complex strains. *PeerJ* 11: 1-13 DOI:10.7717/peerj.5895
534 Krishnan N, Malaga W, Constant P, Caws M, Tran TH, Salmons J, Nguyen TN, Nguyen DB,
535   Daffé M, Young DB, Robertson BD, Guilhot C, Thwaites GE. 2011. *Mycobacterium*
536   *tuberculosis* lineage influences innate immune response and virulence and is associated
537   with distinct cell envelope lipid profiles. *PLoS ONE* 6: 1-6 DOI:
538   10.1371/journal.pone.0023870.
539 Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and
540   annotation of phylogenetic and other trees. *Nucleic Acids Research* 44: W242–245
541   DOI:10.1093/nar/gkw290
542 Malik AN, Godfrey-Faussett P. 2005. Effects of genetic variability of *Mycobacterium*
543   *tuberculosis* strains on the presentation of disease. *Lancet Infectious Disease* 5: 174–83
544   DOI: 10.1016/S1473-3099(05)01310-1
545 McHenry ML, Bartlett J, Igo RP, Wampande EM, Benchek P, Mayanja-Kizza H, Benchek P,
546   Mayanja-Kizza H, Fluegge K, Hall NB, Gagneux S, Tishkoff SA, Wejse C, Sirugo G,
547   Boom WH, Joloba M, Williams SM, Stein CM. 2020. Interaction between host genes and
548   *Mycobacterium tuberculosis* lineage can affect tuberculosis severity: Evidence for
549   coevolution?. *PLoS genetics*, 16: 1-18. DOI: 10.1371/journal.pgen.1008728.
550 Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MG, Rüsch-Gerdes
551   S, Mokrousov I, Aleksic E, Allix-Béguec C, Antierens A, Augustynowicz-Kopeć E,
552   Ballif M, Barletta F, Beck HP, Barry CE 3rd, Bonnet M, Borroni E, Campos-Herrero I,
553   Cirillo D, Cox H, Crowe S, Crudu V, Diel R, Drobniewski F, Fauville-Dufaux M,
554   Gagneux S, Ghebremichael S, Hanekom M, Hoffner S, Jiao WW, Kalon S, Kohl TA,
555   Kontsevaya I, Lillebæk T, Maeda S, Nikolayevskyy V, Rasmussen M, Rastogi N, Samper
556   S, Sanchez-Padilla E, Savic B, Shamputa IC, Shen A, Sng LH, Stakenas P, Toit K,
557   Varaine F, Vukovic D, Wahl C, Warren R, Supply P, Niemann S, Wirth T. 2015.
558   Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing
559   lineage. *Nature Genetics* 47: 242–249 DOI:10.1038/ng.3195
560 Mokrousov I, Ly HM, Otten T, Lan NN, Vyshnevskyi B, Hoffner S, Narvskaya O. 2005. Origin
561   and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from
562   human phylogeography. *Genome Research* 15: 1357–1364 DOI:10.1101/gr.3840605
563 Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, Antoine R, Niyigena
564   EB, Mulders W, Fissette K, Diels M, Gaudin C, Duthoy S, Ssengooba W, André E,
565   Kaswa MK, Habimana YM, Brites D, Affolabi D, Mazarati JB, de Jong BC, Rigouts L,
566   Gagneux S, Meehan CJ, Supply P. 2020. A sister lineage of the *Mycobacterium*
567   *tuberculosis* complex discovered in the African Great Lakes region. *Nature*
568   *communications* 11: 1-11 DOI:10.1038/s41467-020-16626-6
569 Nguyen L. 2016. Antibiotic resistance mechanisms in *M. tuberculosis*: an update. *Archives of*
570   *Toxicology* 90:1585–1604 DOI:10.1007/s00204-016-1727-6
571 Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective
572   stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology*
573   *and Evolution* 32: 268–274 DOI:10.1093/molbev/msu300

574 Portevin D, Gagneux S, Comas I, Young D. 2011. Human macrophage responses to clinical
575   strains from the *Mycobacterium tuberculosis* complex discriminate between ancient and
576   modern lineages. *PLoS Pathogens* 7: 1-12 DOI:10.1371/journal.ppat.1001307
577 Rakotosamimanana N, Raharimanga V, Andriamandimby SF, Soares JL, Doherty TM,
578   Ratsitorahina M, Ramarokoto H, Zumla A, Huggett J, Rook G, Richard V, Gicquel B,
579   Rasolofo-Razanamparany V; VACSEL/VACSIS Study Group. 2010. Variation in gamma
580   interferon responses to different infecting strains of *Mycobacterium tuberculosis* in acid-
581   fast bacillus smear-positive patients and household contacts in Antananarivo,
582   Madagascar. *Clinical Vaccine Immunology* 17: 1094–1203 DOI:10.1128/CVI.00049-10
583 Rodríguez-Castillo JG, Pino C, Niño LF, Rozo JC, Llerena-Polo C, Parra-López CA, Tauch A,
584   Murcia-Aranguren MI. 2017. Comparative genomic analysis of *Mycobacterium*
585   *tuberculosis* Beijing-like strains revealed specific genetic variations associated with
586   virulence and drug resistance. *Infection Genetics and Evolution* 54: 314–323
587   DOI:10.1016/j.meegid.2017.07.022
588 Sarkar R, Lenders L, Wilkinson KA, Wilkinson RJ, Nicol MP. 2012. Modern lineages of
589   *Mycobacterium tuberculosis* exhibit lineage-specific patterns of growth and cytokine
590   induction in human monocyte-derived macrophages. *PLoS ONE* 7: 1-8
591   DOI:10.1371/journal.pone.0043170
592 Shitikov E, Kolchenko S, Mokrousov I, Bespyatykh J, Ischenko D, Ilina E, Govorun V. 2017.
593   Evolutionary pathway analysis and unified classification of East Asian lineage of
594   *Mycobacterium tuberculosis*. *Scientific Reports* 7: 1-10 DOI:10.1038/s41598-017-10018-
595   5
596 Singh J, Sankar MM, Kumar P, Couvin D, Rastogi N, Singh S, Indian TB Diagnostics Network.
597   2015. Genetic diversity and drug susceptibility profile of *Mycobacterium tuberculosis*
598   isolated from different regions of India. *Journal of Infectology* 71: 207–219
599   DOI:10.1016/j.jinf.2015.04.028
600 Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S,
601   Luo T, Gao Q, Kato-Maeda M, Ballif M, Egger M, Macedo R, Mardassi H, Moreno M,
602   Tudo Vilanova G, Fyfe J, Globan M, Thomas J, Jamieson F, Guthrie JL, Asante-Poku A,
603   Yeboah-Manu D, Wampande E, Ssengooba W, Joloba M, Henry Boom W, Basu I,
604   Bower J, Saraiva M, Vaconcellos SEG, Suffys P, Koch A, Wilkinson R, Gail-Bekker L,
605   Malla B, Ley SD, Beck HP, de Jong BC, Toit K, Sanchez-Padilla E, Bonnet M, Gil-
606   Brusola A, Frank M, Penlap Beng VN, Eisenach K, Alani I, Wangui Ndung'u P, Revathi
607   G, Gehre F, Akter S, Ntoumi F, Stewart-Isherwood L, Ntinginya NE, Rachow A,
608   Hoelscher M, Cirillo DM, Skenders G, Hoffner S, Bakonyte D, Stakenas P, Diel R,
609   Crudu V, Moldovan O, Al-Hajoj S, Otero L, Barletta F, Jane Carter E, Diero L, Supply P,
610   Comas I, Niemann S, Gagneux S. 2016. *Mycobacterium tuberculosis* lineage 4 comprises
611   globally distributed and geographically restricted sublineages. *Nature Genetics* 48: 1535-
612   1543 DOI:10.1038/ng.3704.
613 Svensson E, Millet J, Lindqvist A, Olsson M, Ridell M, Rastogi N; Western Sweden
614   Tuberculosis Epidemiology Study Group. 2011. Impact of immigration on tuberculosis
615   epidemiology in a low-incidence country. *Clinical Microbiology and Infection* 17: 881–
616   887 DOI:10.1111/j.1469-0691.2010.03358.x
617 Thwaites G, Caws M, Chau TT, D'Sa A, Lan NT, Huyen MN, Gagneux S, Anh PT, Tho DQ,
618   Torok E, Nhu NT, Duyen NT, Duy PM, Richenberg J, Simmons C, Hien TT, Farrar J.
619   2018. Relationship between *Mycobacterium tuberculosis* genotype and the clinical

620   phenotype of pulmonary and meningeal tuberculosis. *Journal of Clinical Microbiology*
621   46: 1363–1368 DOI:10.1128/JCM.02180-07
622   Wang C, Peyron P, Mestre O, Kaplan G, van Soolingen D, Gao Q, Gicquel B, Neyrolles O.
623   2010. Innate immune response to *Mycobacterium tuberculosis* Beijing and other
624   genotypes. *PLoS ONE* 5: 1-8 DOI:10.1371/journal.pone.0013594
625   Wang XH, Ma AG, Han XX, Gu XM, Fu LP, Li PG, Li FY, Wang QZ, Liang H, Katar A, Wang
626   LJ. 2015. Correlations between drug resistance of Beijing/W lineage clinical strains of
627   *Mycobacterium tuberculosis* and sublineages: a 2009–2013 prospective study in Xinjiang
628   province, China. *Medical Science Monitor* 21: 1313–1318 DOI:10.12659/MSM.892951
629   Xia E, Teo YY, Ong RT. 2016. SpoTyping: Fast and accurate in silico Mycobacterium
630   spoligotyping from sequence reads. *Genome Medicine* 8: 1–9 DOI:10.1186/s13073-016-
631   0270-7

Figure 1

Phylogenetic analysis of *M. tuberculosis strains* from PTB and EPTB phenotypes of the disease. Sensitive, does not present genotypic resistance; Drug-resistant, resistant to at least one antibiotic; MDR, resistant to at least isoniazid and rifampicin

Sensitive, does not present genotypic resistance; Drug-resistant, resistant to at least one antibiotic; MDR, resistant to at least isoniazid and rifampicin; XDR, resistant to isoniazid and rifampicin plus any fluoroquinolone and at least one of three injectable second-line drugs. The phylogenetic tree was inferred using the maximum likelihood (ML) criterion with a general time-reversible model of nucleotide substitution and a gamma model of rate heterogeneity. Yellow highlighted letters indicate EPTB strains. Support values correspond to bootstrap values. The topology was rooted with a *Mycobacterium microti* strain.
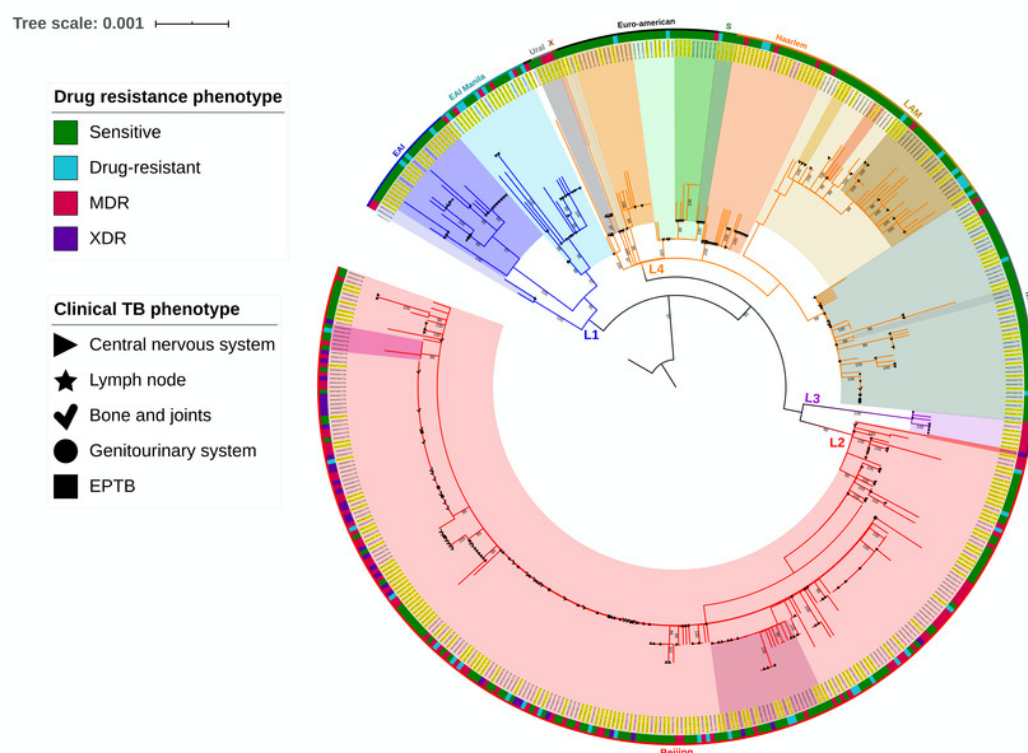
Figure 2

Phylogeny of 263 Mycobacterium tuberculosis L2 strains. The phylogeny was constructed by the maximum likelihood (ML) criterion. Classification of strains into sublineages and informative genetic markers are shown. Support values correspond to bootstrap va

The phylogeny was constructed by the maximum likelihood (ML) criterion. Classification of strains into sublineages and informative genetic markers are shown. Support values correspond to bootstrap values