Whole-genome comparative analysis at the lineage/sublineage level disclose relationships between *Mycobacterium tuberculosis* genotype with 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 extrapulmonary (EPTB) phenotype affecting any organ or tissue. MTBC species includes nine phylogenetic lineages, some appearing globally, and others that are geographically restricted; a high incidence of infections associated with the former suggests it to be more virulent than the latter. 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 with PTB/EPTB phenotypes, have shown inconsistent results attributed to the strains predominance from specific *M. tuberculosis* lineage/sublineage in the sample analyzed, and because of the use of low-resolution phylogenetic tools which have impaired the sublineages discrimination. Present work elucidates the relationships between MTBC strain lineage/sublineage with the clinical phenotype of the disease and its antibiotic resistance. **Methods**. To avoid biases, we retrieved the raw genomic reads (RGR) of all (n=245) the *M. tuberculosis* strains worldwide causing EPTB available in databases, and an evenly representative sample of RGR (n=245) of PTB strains. With the whole assembled genomes obtained from these RGR, a robust maximum likelihood phylogeny based on single nucleotide polymorphisms was generated, allowing an effective strains lineage/sublineage assignment. **Results.** A significant association (O.R. range: 1.1-8.1) was found between EPTB and ancient sublineages, and between PTB with modern sublineages. 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 the intralineage diversity can drive differences in the immune response that triggers the PTB/EPTB phenotype.

Whole-genome comparative analysis at the lineage/sublineage level disclose relationships between *Mycobacterium tuberculosis* genotype with clinical phenotype.

Andrea Monserrat Negrete-Paz[1], Gerardo Vázquez-Marrufo[1], Ma. Soledad Vázquez-Garcidueñas[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, Mexico

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

Abstract

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 extrapulmonary (EPTB) phenotype affecting any organ or tissue. MTBC species includes nine phylogenetic lineages, some appearing globally, and others that are geographically restricted; a high incidence of infections associated with the former suggests it to be more virulent than the latter. 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 with PTB/EPTB phenotypes, have shown inconsistent results attributed to the strains predominance from specific *M. tuberculosis* lineage/sublineage in the sample analyzed, and because of the use of low-resolution phylogenetic tools which have impaired the sublineages discrimination. Present work elucidates the relationships between MTBC strain lineage/sublineage with the clinical phenotype of the disease and its antibiotic resistance.
Methods. To avoid biases, we retrieved the raw genomic reads (RGR) of all (n=245) the *M. tuberculosis* strains worldwide causing EPTB available in databases, and an evenly representative sample of RGR (n=245) of PTB strains. With the whole assembled genomes

obtained from these RGR, a robust maximum likelihood phylogeny based on single nucleotide polymorphisms was generated, allowing an effective strains lineage/sublineage assignment. **Results.** A significant association (O.R. range: 1.1-8.1) was found between EPTB and ancient sublineages, and between PTB with modern sublineages. 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 the intralineage diversity can drive differences in the immune response that triggers the PTB/EPTB phenotype.

Introduction

Tuberculosis (TB) represents the main cause of death among infectious diseases worldwide, with its drug-resistant manifestations constituting a major global health concern (Floyd *et al*., 2018). Pulmonary TB (PTB) is the most common disease clinical phenotype, but some patients develop extrapulmonary (EPTB) phenotype affecting practically any organ or tissue, including the aggressive manifestations of lymph node and central nervous system TB (Golden & Vikram, 2005). Human TB is caused by members of the *Mycobacterium tuberculosis* complex (MTBC), with >99% nucleotide sequence identity at genomic level (Gagneux, 2018). The human-adapted species of the MTBC are *M. tuberculosis sensu stricto* and *Mycobacterium africanum*, which are divided into nine phylogenetic lineages: L1 or Indo-Oceanic, L2 or East Asian, L3 or East African-Indian, L4 or Euro-American, L5 or *M. africanum* West-African 1, L6 or *M. africanum* West-African 2, L7 or Ethiopia (Firdessa *et al.,* 2013; Coscolla & Gagneux, 2014), L8 or *M. africanum* from African Great Lakes *(*Ngabonziza *et al.,* 2020*)* and the recently described *M. africanum* L9 (Coscolla *et al.,* 2021). Global phylogeographic reconstruction of *M. tuberculosis* suggests that each lineage has become specifically adapted to defined human populations (Gagneux, 2018), with some occurring globally and others showing a strong geographical restriction. High incidence of infection cases associated with non-geographically restricted strains might imply that these strains are more effective in causing the disease (Malik & Godfrey-Fausset, 2005).

Several studies have been conducted to identify possible relationships between *M. tuberculosis* phylogenetic lineages with the PTB or EPTB phenotype of the disease (Feng *et al*., 2008; Click *et al*., 2012), but results show a lack of consistency. Different factors contribute to explaining the observed discrepancies among the conducted studies in an attempt to establish genotype-phenotype relationships. In the first instance, there are differences in the sample size (Coscolla & Gagneux, 2014), as well as in the non-homogeneous distribution of lineages among the set (Kato-Maeda & Nahid, 2012) of analyzed strains. Moreover, biased associations might arise due to a lack of data or failure to control for possible confounders associated with known risk factors for EPTB, such as HIV comorbidity in patients from whom strains were isolated (Coscolla & Gagneux, 2014). Furthermore, the studies use different operational definitions for EPTB (Kato-Maeda & Nahid, 2012), and some lack appropriate tools to index genomic diversity and classify strains into lineages of some studies (Coscolla & Gagneux, 2010). Whole-genome

comparative analysis has allowed categorizing *M. tuberculosis* lineages into sublineages using single nucleotide polymorphism (SNP) analysis (Coll *et al*., 2014; Stucki *et al*., 2016)) This subtle level of strain differentiation suggests that some sublineages might drive the observed associations of a *M. tuberculosis* genotype with a specific disease phenotype, such as EPTB (Feng *et al*., 2008), but the analysis to prove this hypothesis has not been conducted. Interestingly, the frequency of the sublineages assigned to isolated strains from the East Asian (L2) lineage differs among populations settled in different geographical areas, and that might also explain why some studies associate L2 lineage strains with EPTB (Feng *et al*., 2008), whereas others associate it with PTB (Dale *et al*., 2005), and still others do not find any association at all (Svensson *et al.,* 2011). Thus, intralineage diversity requires a detailed exploration to clarify disease phenotype variation and its relationship with MTBC genotypes (Kato-Maeda & Nahid, 2012). Additionally, genomic evidence has revealed a close relationship among specific *M. tuberculosis* lineages and sublineages with drug resistance (Wang *et al*., 2015), a relevant public health phenotype that might hinder successful TB treatment. Horizontal transfer of drug resistance genes has not been reported in *M. tuberculosis*, but resistance mostly arises from chromosomal mutations under the selective pressure of antibiotic use (Nguyen, 2016). Thus, characterization of mutations associated with resistance phenotypes in strains with different genotypes can help to reveal lineage/sublineage specific microevolutionary processes of epidemiological relevance. Drug resistance-associated mutations have been hypothesized to modify the strain fitness and its ability to cross the blood-brain barrier causing EPTB, specifically tuberculous meningitis (Faksri *et al*., 2018).

Unclarified relationships of lineages/sublineages with *M. tuberculosis* PTB/EPTB disease and drug resistance phenotypes hinder the generation of adequate epidemiological transmission chains and timely successful treatments. Therefore, this work aims to elucidate relationships between strain genotype at the lineage/sublineage level with the disease clinical phenotype and antibiotic resistance. We retrieved the raw reads datasets from all the *M. tuberculosis* strains causing EPTB worldwide deposited in databases to assemble whole-genome sequences and, consequently, avoid sources of possible biases related to previous studies originated by (i) a low number of analyzed strains, (ii) PTB/EPTB unequal strains analyzed, or (iii) differences on regional clinical TB phenotype incidences. Furthermore, the raw datasets from genomes of the same number of strains causing PTB worldwide were included. The epidemiological and public health relevance of such relationships is discussed.

Materials & Methods

Data retrieval

A total of 490 raw datasets of genome sequence reads were retrieved, which corresponded to 245 *M. tuberculosis* strains causing PTB and 245 strains causing EPTB (Supplemental Table 1). We considered EPTB strains were those from cases for which the major site of infection reported in databases was not pulmonary or miliary and, an additional infection site either was not specified

or, was specified but not as pulmonary or miliary (Click *et al*., 2012). The sequence quality of the FASTQ reads was checked using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and subsequently filtered to a Phred score of 30 using TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).

In silico typing

In silico spoligotyping was performed using SpoTyping program (Xia, Teo & Ong, 2016) for next-generation sequencing reads. To determine lineage, the TB INSIGHT (http://tbinsight.cs.rpi.edu) server was used based on found spoligotypes.

Phylogenetic reconstruction

To find SNPs for phylogenetic reconstruction (Homolka *et al*., 2012; Coll *et al*., 2014; Merker *et al*., 2015), the sequencing reads of studied strains were aligned to the reference strain of *M. tuberculosis* H37Rv (accession no. NC_000962.3) using the MTBseq program (Kohl *et al*., 2018). Phylogenetic inferences were conducted with the maximum likelihood (ML) criterion using IQ-TREE package (Nguyen *et al*., 2015) with a general time-reversible (GTR) model of nucleotide substitution and a gamma model of rate heterogeneity. Phylogenetic trees were constructed based on 1,000 bootstrap replicates and their visualization was performed using iTOL (Letunic & Bork, 2016).

In silico determination of drug resistance

Raw FASTQ sequencing files were uploaded in TB Profiler version 0.2.1 (Coll et al., 2015), a tool to determine *in silico* drug resistance. It uses raw sequence data as input, aligns the data to the *M. tuberculosis* H37Rv reference genome, and then compares the identified SNPs and indels to a curated list of 1,325 drug resistance mutations. This package also determines the *M. tuberculosis* lineage based on a 62-SNP barcode. The TB Profiler-predicted resistance mutations were validated using the results of MTBseq, which reports a list of mutations in genes associated with antimicrobial resistance for every processed strain.

Data analysis

Data entry and statistical analyses were performed in SPSS version 16 (SPSS Inc., Illinois, USA). The associations among *M. tuberculosis* genotypes, drug resistance, and clinical phenotypes of TB were estimated using multiple regression model and Fisher's exact test, with 95% confidence intervals to assess significance.

Results

Anatomical site of infection of strains selected for comparative genomics analysis

A total of 490 raw genome datasets of *M. tuberculosis* strains isolated from seven different countries with a negative human immunodeficiency virus infection status (Supplemental Table 1) were genotyped. The selected strains included those from countries where specific lineages

predominate according to SITVIT web (http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE/), such as Thailand and Indonesia for the East Asian lineage and India for the East African-Indian lineage. The EPTB phenotype strains were clustered into five major groups according to the anatomical site of infection, with a predominance of central nervous system (72.24%). Other anatomical sites of the disease were bone and joints (17.14%), lymph nodes (4.49%), and the genitourinary system (2.86%); the remaining 3.27% were from sites just specified as EPTB strains in databases.

Distribution and association of lineages and genotypes with clinical phenotype

The strains were assigned to a major *M. tuberculosis* genetic lineage according to the spoligotype using the TB-lineage search option of the TB-insight web server. These results correlated with the clustering pattern generated by SNP-based phylogeny (Figure 1). The predominant lineage of the analyzed strains was East Asian (54.28%), followed by Euro-American (34.70%), Indo-Oceanic (9.6%), and East African-Indian (1.42%). The EPTB cases percentage differed among these phylogenetic lineages, with 15.51% for Indo-Oceanic, 48.98% for East Asian, 2.45% for East African-Indian, and 33.06% for Euro-American. The same variation in phylogenetic lineages was observed for strains from PTB cases, with 59.59%, 36.32%, 3.67%, and 0.40% for East Asian, Euro-American, Indo-Oceanic, and East African-Indian, respectively. Both Fisher's exact test and multiple logistic regression model showed that PTB was significantly associated with the East-Asian lineage strains (OR 1.5; 95% CI 1.1–2.2) and that EPTB was significantly associated with the Indo-Oceanic lineage strains (OR 4.8; 95% CI 2.2–10.1).

We searched for a relationship among the phylogenetic lineages found in the analyzed strains and anatomical sites of infection, including lungs, lymph nodes, genitourinary system, central nervous system, as well as bone and joints. The results showed that strains belonging to the Indo-Oceanic lineage were significantly associated with central nervous system infection (OR 3.94; CI 95% 2.1-7.4), whereas those of the East Asian lineage were significantly associated with bone and joints infection (OR 1.5; 95% CI 1.3-1.9) but also with PTB strains, as previously described. Further division of identified lineages was conducted using a SNP analysis, distinguishing 27 unique sublineages, with the predominant sublineages being 2.2.1 (n=218 strains), 2.2.1.1 (n=25), 4.1.2.1 (n=22), 4.3.1 (n=24), 4.8 (n=52), and 1.2.1 (n=23). As expected, statistical analysis showed significant relationships of some of these sublineages with major infection sites. In this way, sublineage 1.1.1 was associated with central nervous system infection (OR 2.8; 95% CI 1.0–7.7), and the same relationship was observed for sublineages 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, 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; 95% CI 1.0-13.8) were associated with the pulmonary phenotype.

A deeper whole-genome SNP-based classification to discriminate the proto-Beijing from the Beijing strains allowed us to identify the outbreak sublineages of Asia Ancestral 1, Asia Ancestral 3, Asian African 2, Asian African 2/RD142, Asian African 3, Pacific RD150, Central Asia, and Europe/Russia B0/W148 and a group of modern unclassified strains using informative

genetic markers found in the genomes of studied strains (Figure 2). Of all these genotypes, the Asian African 2 sublineage (OR 2.3; 95% CI 1.1-5.2) and Europe/Russia B0/W148 outbreak sublineage (OR 2.7; 95% CI 1.3-5.4) were significantly associated with PTB. In the same way, the Ancestral Beijing sublineages (OR 2.4; 95% CI 1.2-4.8) and Central Asia subgroup (OR 8.2; 95% CI 2.9-22.9) were significantly associated with EPTB.

The most frequent spoligotype among L2 strains (85.1%), was SIT1 (Table S2). The SIT 53 and SIT 19 were the most frequent spoligotypes for L4 (22.9%) and L1 (27.6%) strains, respectively. For L3 strains, 42% had an unclassified SIT, spoligotype (703775740003771). Spoligotyping allowed identification of twelve major genotypic families, including Beijing (53.7%), Central Asia (CAS) (1.4%), EAI (9.8%), Haarlem (4.4%), Latin American–Mediterranean (LAM) (7.5%), S (0.2%), T (19.8%), X (0.4%) and Family33-36 (2%). Interestingly, 13.26% of the strains corresponded to spoligotypes not previously reported in the SITVIT database (Supplemental Table 2). A significant association was found between the family EAI2 (OR 6.7; 95% CI 2.0-22.1) and EPTB, of which 13 strains caused tuberculous meningitis and two of them led to lymph nodes infection.

Mutations associated with drug-resistant TB

A total of 78 previously reported mutations distributed in 17 genes known to confer resistance to first and second-line drugs for TB treatment were identified (Supplemental Table 3). The most common mutations associated with rifampin resistance were found in genes *rpoB* and *rpoC* in 36.93% of the studied strains; 38.36% had mutations in *katG* and *inhA* genes, associated with isoniazid resistance, and 28.36% had mutations in the *rpsL* gene, associated with streptomycin resistance. Extrapulmonary strains showed greater diversity of mutations but appeared less frequently than pulmonary strains. Mutations *gyrA* (Ala90Val) and *tlyA* (Lys69Glu) were found only in pulmonary strains. On the other hand, mutations in genes *rpoB* (Glu761Asp, Leu430Pro, Leu464Met), *folC* (Glu153Ala), *rpsL* (Lys88Thr), *rrs* (514a>c), *embA* (16C>G), *embB* (Ala356Val), *inhA* (Ile21Thr), *katG* 589_590insGT, and *pncA* (Ala3Glu, Cys138Arg, Cys72Arg, Gln10Pro, His51Asp, Leu159Arg, Pro54Leu, Thr135Pro) were only found in extrapulmonary strains. There were no statistically significant differences among these variations with the tuberculosis clinical phenotype.

A similar distribution of drug-resistant (5.9% PTB, 4.2% EPTB) and sensitive (26.1% PTB, 35.5% EPTB) strains were observed in EPTB and PTB (Supplementary Table 3). 16.4% of central nervous system and 82.5% of bone and joint strains showed resistance to isoniazid, and 81.8% of lymph node strains were resistant to rifampin. Multidrug-resistance (MDR; OR:2.1; 95% IC 1.4-3.3) and extensively drug-resistance (XDR; OR:8.1; 95% IC 2.8-13.4) were strongly associated with PTB strains than with EPTB strains. We also observed a significant association of Lineage 3 strains with MDR (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).

Discussion

In this work, we used both raw genome datasets and whole-genome SNP-based phylogeny to genotype the same number of PTB and EPTB *M. tuberculosis* strains. All the strains were included in one of the six lineages defined by Coll (Coll *et al*., 2014). When seeking for lineage-clinical phenotype relationships, we found that the East-Asian 'modern' *M. tuberculosis* lineages were associated with PTB, whereas the Indo-Oceanic so-called 'ancestral lineages' were associated with EPTB. It is known that strains from 'modern' lineages induced a slighter inflammatory response than those from 'ancient' lineages, which has been related to a selective advantage of strains from 'modern' lineages, resulting in impaired bacterial control by the host, a faster disease progress, and enhanced transmission (Portevin *et al*., 2011). These clinical differences contribute to explain the association of the East-Asian lineage with the PTB phenotype, supporting the hypothesis that since this is the disease contagious phenotype, patients infected with strains of such a lineage are prone to develop the pulmonary disease at a higher frequency than those patients infected with strains from other genotypes, which is consistent with its increased transmissibility (Thwaites *et al*., 2018).

It has been proposed that strains belonging to the Indo-Oceanic lineage are 'less virulent' than those from other lineages and cause a specific exacerbated inflammatory response (Chakraborty *et al*., 2018), which might be attributed to the presence of unique cell envelope lipids in this lineage, such as phenolphthiocerol dimycocerosate (Krishnan *et al*., 2011). Nevertheless, it is still unknown if these differences in mycobacterial cell envelope lipid composition can explain lineage-related phenotypic differences such as the EPTB phenotype. Interestingly, the percentage of EPTB cases associated with the East African-Indian lineage was as high as 85%, but it did not show statistical significance due to the small number of EPTB raw genomes (n=7) from this lineage available in databases. In this regard, the present work reveals databases gaps relevant for its public health and epidemiological implication, i.e., the need to include more East African Indian lineage genomes to clarify its relationships with EPTB.

On the search for a detailed relationship between phylogenetic lineages and the anatomical site of infection, strains belonging to the East Asian lineage were associated with infection both of bone and joints and PTB. East Asian lineage comprises two major clades or sublineages, designated as proto-Beijing (2.1) and Beijing (2.2) (Ajawatanawong *et al*., 2019). Sublineage 2.2 or the Beijing family, as defined by spoligotyping, is composed of several sublineages broadly categorized into the Ancestral and Modern Beijing strains (Mokrousov *et al*., 2005). A subtle SNP-based classification was recently proposed that allows the discrimination of the proto-Beijing from the Beijing strains (Shitikov *et al*., 2017). Such classification divides the Beijing group into the Ancestral Beijing clade and the Modern Beijing clade, which comprises two groups, one including three strains (Asia Ancestral 1, Asia Ancestral 2, Asia Ancestral 3) and the other including seven strains (Asian African 1, Asian African 2, Asian African 2/RD142, Asian African 3, Pacific RD150, Europe/Russia B0/W148 outbreak and Central Asia), respectively. This classification allows us to associate the Asian African 2 sublineage and Europe/Russia B0/W148 outbreak sublineage with PTB, as well as the Ancestral

Beijing sublineages and Central Asia subgroup with EPTB. Differences in pathogenicity of Beijing sublineages have been previously reported (Feng *et al.,* 2008), and sublineage specific patterns of induced cytokine production by macrophages have also been observed (Sarkar *et al*., 2012). In this regard, a macrophage infection model revealed that the Ancestral Beijing strains induce a higher production of pro-inflammatory cytokines TNF-α and IL-6 than the Modern Beijing sublineage (Chen *et al*., 2014). High IFN-γ expression and cytokine production have also been reported in peripheral blood mononuclear cells of Ancestral Beijing strains (Faksri *et al*., 2014). Such results suggest that the Ancestral Beijing strains are as highly immunogenic (Kato-Maeda *et al*., 2012) as Lineage 1 strains (EAI) (Rakotosamimanana *et al*., 2010). Interestingly, these two sublineages (EAI and Ancestral Beijing) were associated with the EPTB phenotype in this study. In contrast, the other sublineage associated with EPTB in this work 4.1.2.1, has been reported to induce similar cytokine levels to H37Rv (Haarlem family strains), which are associated with a low immune response (Wang *et al*., 2010). This supports the idea that strains of different sublineages vary by many phenotypes such as the tendency to develop drug resistance, virulence levels, and immune response, which influence the disease severity and clinical presentation.

Interestingly, the sublineages 4.1.2.1 and Ancestral Beijing belonging to Lineages 4 and 2, respectively, show a high prevalence worldwide, representing more than 50% of the strains in certain areas and/or subpopulations (Ajawatanawong *et al*., 2019). In contrast, Lineage 1 (EAI) strains are commonly reported in countries around the Indian Ocean, being one of the most geographically restricted families. However, EAI strains have been reported in lower percentages in countries such as the Netherlands, Australia, the USA, Sweden, Saudi Arabia, Tunisia, Taiwan, Panama, and Mexico. The East-Asian India 2 spoligotype of Lineage 1 corresponds to the Nonthaburi (EAI2-Nonthaburi) genotype and the Manila (EAI2-Manila) genotype (Couvin, Reynaud & Rastogi, 2019). Recently, Coker (Coker *et al*. 2016) reported a polymorphism in the genome of Nonthaburi strains from three patients with tuberculous meningitis, consisting of a 500 bp deletion covering ppe50, that was not present in the reference strain *M. tuberculosis* H37Rv. They reported three mutations (T28910C, C1180580T and, C152178T) until now found only in these meningeal Nonthaburi strains. These mutations could represent part of the genetic background that could be shared by strains that cause EPTB and that also belong to the EAI, 4.1.2.1, and Ancestral Beijing sublineages. Nonetheless, further investigation is required to determine if these mutations are shared, which could be the functional consequences and the probable relation of such polymorphisms with TB disease phenotype.

Regarding genotype relationship with antibiotic resistance, to the best of our knowledge, a significant association of Lineage 3 with drug resistance found here has not been previously reported. This result might be useful to optimize TB treatment in geographical areas where this lineage is frequent. Studies from Asia, Europe and Africa have shown varying associations between drug resistance and MTB lineages (Coscolla & Gagneux, 2010; Singh et al., 2015); however, as was found here, Beijing strains have been associated with MDR and XDR in several cases (Rodríguez-Castillo *et al*., 2017).

The advantages of the present study regarding previous works seeking for relationships of *M. tuberculosis* genotype with clinical phenotype include the use of a significantly higher number of strains, the same number of EPTB and PTB strains, a clear assignment of the strains considered as EPTB, and exclusion of the strains from HIV patients, being this latter a controlled variable, a caution not commonly considered in similar works. However, other possible confounders or known risk factors for EPTB could not be considered in the analysis here conducted, partly due to the lack of additional metadata of clinical information in databases where the genomes of the strains were retrieved. This lack of metadata is not a source of bias in the relationships found here because it has been shown that the association between lineage and EPTB found among the US-born population remains similar to that of the overall population after filtering by race/ethnicity, HIV infection status, age, and sex (Click *et al.,* 2012). These data support the assertion that the relationship between lineage and the disease clinical phenotype is not attributable to regional-specific factors, so if this association exists, it does not depend on these unknown clinical factors.

Conclusions

Overall, the obtained results suggest that intralineage diversity could drive differences in the immune response that trigger the different clinical phenotypes of tuberculosis. The immune response caused by Ancestral lineages (L1) and Ancestral Beijing strains could be similar, eliciting a high inflammatory response. Our results highlight the need to analyze both the genomic background shared by these strains and to perform in vitro/in vivo studies that help to elucidate the mechanism by which they could disseminate through the body, causing extrapulmonary disease. We demonstrated that the lack of consistency regarding clinical phenotype-strain genotype in the results obtained by previous studies is partially due to the use of inadequate phylogenetic classification tools, which do not allow discrimination between sublineages. Additionally, the present results were not biased by the predominance of a specific lineage/sublineage in a specific human population because we included all the EPTB strains available in databases, and an evenly representative sample of PTB strains.

Author contributions

A.M.N.P. and G.V.M. contributed equally to this article.

References

1. Floyd K, Glaziou P, Zumla A, Raviglione M. 2018. The global tuberculosis epidemic progress in patient care, prevention and control efforts in year 3 of the End TB Era. *Lancet Respiratory Medicine* 6: 299–314 DOI:10.1016/S2213-2600(18)30057-2
2. Golden MP, Vikram HR. 2005. Extrapulmonary tuberculosis: an overview. *American Family Physician* 72: 1761–1768.

3. Gagneux S. 2018. Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews Mycrobiology* 16: 202–213 DOI: 10.1038/nrmicro.2018.8
4. Coscolla M, Gagneux S. 2014. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Seminars in Immunology* 26: 431–44 DOI:10.1016/j.smim.2014.09.012
5. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, Gadisa E, Kiros T, Habtamu M, Hussein J, Zinsstag J, Robertson BD, Ameni G, Lohan AJ, Loftus B, Comas I, Gagneux S, Tschopp R, Yamuah L, Hewinson G, Gordon SV, Young DB, Aseffa A. 2013. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerging Infectious Disease* 19: 460-463 DOI:10.3201/eid1903.120256.
6. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F , Tzfadia O, Antoine R, Niyigena EB, Mulders W, Fissette K, Diels M, Gaudin C, Duthoy S, Ssengooba W, André E, Kaswa MK, Habimana YM, Brites D, Affolabi D, Mazarati JB, de Jong BC, Rigouts L, Gagneux S, Meehan CJ, Supply P. 2020. A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region. *Nature communications* 11: 1-11 DOI:10.1038/s41467-020-16626-6
7. Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, Borrell S, Otchere ID, Asante-Poku A, Asare P, Sánchez-Busó L, Gehre F, Sanoussi CN, Antonio M, Affolabi D, Fyfe J, Beckert P, Niemann S, Alabi AS, Grobusch MP, Kobbe R, Parkhill J, Beisel C, Fenner L, Böttger EC, Meehan CJ, Harris SR, de Jong BC, Yeboah-Manu D, Brites D. 2021. Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. *Microbial Genomics* 7:1-14 DOI:10.1099/mgen.0.000477
8. Malik AN, Godfrey-Faussett P. 2005. Effects of genetic variability of *Mycobacterium tuberculosis* strains on the presentation of disease. *Lancet Infectious Disease* 5: 174–83 DOI: 10.1016/S1473-3099(05)01310-1
9. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. 2012. Relationship between *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis. *Clinical Infectious Disease* 54: 211–219 DOI:10.1093/cid/cir788
10. Feng JY, Su WJ, Tsai CC, Chang SC. 2008. Clinical impact of *Mycobacterium tuberculosis* W-Beijing genotype strain infection on aged patients in Taiwan. *Journal of Clinical Microbiology* 46: 3127–3129 DOI:10.1128/JCM.01132-08
11. Kato-Maeda M, Nahid P. 2012. *Mycobacterium tuberculosis* lineage-what's in your lungs*?*. *Clinical Infectious Disease* 54: 220–224 DOI:10.1093/cid/cir795
12. Coscolla M, Gagneux S. 2010. Does *M. tuberculosis genomic* diversity explain disease diversity?. *Drug Discovery Today Disease Mechanisms* 7: 1–26 DOI:10.1016/j.ddmec.2010.09.004
13. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. 2014. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nature communications* 5: 1-5 DOI:10.1038/ncomms5812
14. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T, Gao Q, Kato-Maeda M, Ballif M, Egger M, Macedo R, Mardassi H, Moreno M, Tudo Vilanova G, Fyfe J, Globan M, Thomas J, Jamieson F, Guthrie JL, Asante-Poku A, Yeboah-Manu D, Wampande E, Ssengooba W, Joloba M, Henry Boom W, Basu I, Bower J, Saraiva M, Vaconcellos SEG, Suffys P, Koch A, Wilkinson R, Gail-Bekker L, Malla B, Ley SD, Beck HP, de Jong BC, Toit K, Sanchez-Padilla E, Bonnet M, Gil-Brusola A, Frank M, Penlap Beng VN, Eisenach K, Alani I, Wangui Ndung'u P, Revathi G, Gehre F, Akter S, Ntoumi F, Stewart-Isherwood L, Ntinginya NE, Rachow A,

Hoelscher M, Cirillo DM, Skenders G, Hoffner S, Bakonyte D, Stakenas P, Diel R, Crudu V, Moldovan O, Al-Hajoj S, Otero L, Barletta F, Jane Carter E, Diero L, Supply P, Comas I, Niemann S, Gagneux S. 2016. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nature Genetics* 48: 1535-1543 DOI:10.1038/ng.3704.

15. Dale JW, Bothamley GH, Drobniewski F, Gillespie SH, McHugh TD, Pitman R. 2005. Origins and properties of *Mycobacterium tuberculosis* strains in London. *Journal of Medical Microbiology* 54: 575–582 DOI:10.1099/jmm.0.45959-0
16. Svensson E, Millet J, Lindqvist A, Olsson M, Ridell M, Rastogi N; Western Sweden Tuberculosis Epidemiology Study Group. 2011. Impact of immigration on tuberculosis epidemiology in a low-incidence country. *Clinical Microbiology and Infection* 17: 881–887 DOI:10.1111/j.1469-0691.2010.03358.x
17. Nguyen L. 2016. Antibiotic resistance mechanisms in *M. tuberculosis*: an update. *Archives of Toxicology* 90:1585–1604 DOI:10.1007/s00204-016-1727-6
18. Wang XH, Ma AG, Han XX, Gu XM, Fu LP, Li PG, Li FY, Wang QZ, Liang H, Katar A, Wang LJ. 2015. Correlations between drug resistance of Beijing/W lineage clinical strains of *Mycobacterium tuberculosis* and sublineages: a 2009–2013 prospective study in Xinjiang province, China. *Medical Science Monitor* 21: 1313–1318 DOI:10.12659/MSM.892951
19. Faksri K, Xia E, Ong RT, Tan JH, Nonghanphithak D, Makhao N, Thamnongdee N, Thanormchat A, Phurattanakornkul A, Rattanarangsee S, Ratanajaraya C, Suriyaphol P, Prammananan T, Teo YY, Chaiprasert A**.** 2018. Comparative whole-genome sequence analysis of *Mycobacterium tuberculosis* isolated from tuberculous meningitis and pulmonary tuberculosis patients. *Scientific Reports* 8: 1–10 DOI:10.1038/s41598-018-23337-y
20. Xia E, Teo YY, Ong RT. 2016. SpoTyping: Fast and accurate in silico Mycobacterium spoligotyping from sequence reads. *Genome Medicine* 8: 1–9 DOI:10.1186/s13073-016-0270-7
21. Homolka S, Projahn M, Feuerriegel S, Ubben T, Diel R, Nübel U, Niemann S. 2012. High resolution discrimination of clinical *Mycobacterium tuberculosis* complex strains based on single nucleotide polymorphisms. *PLoS ONE* 7:1-11 DOI:10.1371/journal.pone.0039855
22. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MG, Rüsch-Gerdes S, Mokrousov I, Aleksic E, Allix-Béguec C, Antierens A, Augustynowicz-Kopeć E, Ballif M, Barletta F, Beck HP, Barry CE 3rd, Bonnet M, Borroni E, Campos-Herrero I, Cirillo D, Cox H, Crowe S, Crudu V, Diel R, Drobniewski F, Fauville-Dufaux M, Gagneux S, Ghebremichael S, Hanekom M, Hoffner S, Jiao WW, Kalon S, Kohl TA, Kontsevaya I, Lillebæk T, Maeda S, Nikolayevskyy V, Rasmussen M, Rastogi N, Samper S, Sanchez-Padilla E, Savic B, Shamputa IC, Shen A, Sng LH, Stakenas P, Toit K, Varaine F, Vukovic D, Wahl C, Warren R, Supply P, Niemann S, Wirth T. 2015. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nature Genetics* 47: 242–249 DOI:10.1038/ng.3195
23. Kohl TA, Utpatel C, Schleusener V, De Filippo MR, Beckert P, Cirillo DM, Niemann S. 2018. MTBseq: A comprehensive pipeline for whole genome sequence analysis of *Mycobacterium tuberculosis* complex strains. PeerJ 11: 1-13 DOI:10.7717/peerj.5895

24. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32: 268–274 DOI:10.1093/molbev/msu300
25. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* 44: W242–245 DOI:10.1093/nar/gkw290
26. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, Mallard K, Nair M, Miranda A, Alves A, Perdigão J, Viveiros M, Portugal I, Hasan Z, Hasan R, Glynn JR, Martin N, Pain A, Clark TG. 2015. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Medicine* 7: 51–61 DOI:10.1186/s13073-015-0164-0
27. Portevin D, Gagneux S, Comas I, Young D. 2011. Human macrophage responses to clinical strains from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS Pathogens* 7: 1-12 DOI:10.1371/journal.ppat.1001307
28. Thwaites G, Caws M, Chau TT, D'Sa A, Lan NT, Huyen MN, Gagneux S, Anh PT, Tho DQ, Torok E, Nhu NT, Duyen NT, Duy PM, Richenberg J, Simmons C, Hien TT, Farrar J. 2018. Relationship between *Mycobacterium tuberculosis* genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. *Journal of Clinical Microbiology* 46: 1363–1368 DOI:10.1128/JCM.02180-07
29. Chakraborty P, Kulkarni S, Rajan R, Sainis K. 2018. *Mycobacterium tuberculosis* strains from ancient and modern lineages induce distinct patterns of immune responses. *Journal of Infection in Developing Countries* 11: 904–911 DOI:10.3855/jidc.8596
30. Krishnan N, Malaga W, Constant P, Caws M, Tran TH, Salmons J, Nguyen TN, Nguyen DB, Daffé M, Young DB, Robertson BD, Guilhot C, Thwaites GE. 2011. *Mycobacterium tuberculosis* lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. *PLoS ONE* 6: 1-6 DOI: 10.1371/journal.pone.0023870.
31. Ajawatanawong P, Yanai H, Smittipat N, Disratthakit A, Yamada N, Miyahara R, Nedsuwan S, Imasanguan W, Kantipong P, Chaiyasirinroje B, Wongyai J, Plitphonganphim S, Tantivitayakul P, Phelan J, Parkhill J, Clark TG, Hibberd ML, Ruangchai W, Palittapongarnpim P, Juthayothin T, Thawornwattana Y, Viratyosin W, Tongsima S, Mahasirimongkol S, Tokunaga K, Palittapongarnpim P. 2019. A novel Ancestral Beijing sublineage of *Mycobacterium tuberculosis* suggests the transition site to Modern Beijing sublineages. *Scientific Reports* 9: 1-12 DOI:10.1038/s41598-019-50078-3
32. Mokrousov I, Ly HM, Otten T, Lan NN, Vyshnevskyi B, Hoffner S, Narvskaya O. 2005. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. *Genome Research* 15: 1357–1364 DOI:10.1101/gr.3840605
33. Shitikov E, Kolchenko S, Mokrousov I, Bespyatykh J, Ischenko D, Ilina E, Govorun V. 2017. Evolutionary pathway analysis and unified classification of East Asian lineage of *Mycobacterium tuberculosis*. *Scientific Reports* 7: 1-10 DOI:10.1038/s41598-017-10018-5
34. Sarkar R, Lenders L, Wilkinson KA, Wilkinson RJ, Nicol MP. 2012. Modern lineages of *Mycobacterium tuberculosis* exhibit lineage-specific patterns of growth and cytokine

induction in human monocyte-derived macrophages. *PLoS ONE* 7: 1-8 DOI:10.1371/journal.pone.0043170

35. Chen YY, Chang JR, Huang WF, Hsu SC, Kuo SC, Sun JR, Dou HY. 2014. The pattern of cytokine production in vitro induced by ancient and modern Beijing *Mycobacterium tuberculosis* strains. *PLoS ONE* 9: 1-7 DOI:10.1371/journal.pone.0094296
36. Faksri K, Chaiprasert A, Pardieu C, Casali N, Palaga T, Prammananan T, Palittapongarnpim P, Prayoonwiwat N, Drobniewski F. 2014. Heterogeneity of phenotypic characteristics of the modern and ancestral Beijing strains of *Mycobacterium tuberculosis*. *Asian Pacific Journal of Allergy and Immunology* 32: 124–132 DOI:10.12932/AP0361.32.2.2013
37. Kato-Maeda M, Shanley CA, Ackart D, Jarlsberg LG, Shang S, Obregon-Henao A, Harton M, Basaraba RJ, Henao-Tamayo M, Barrozo JC, Rose J, Kawamura LM, Coscolla M, Fofanov VY, Koshinsky H, Gagneux S, Hopewell PC, Ordway DJ, Orme IM. 2012. Beijing sublineages of *Mycobacterium tuberculosis* differ in pathogenicity in the guinea pig. *Clinical Vaccine Immunology* 19: 1227–1237 DOI:10.1128/CVI.00250-12
38. Rakotosamimanana N, Raharimanga V, Andriamandimby SF, Soares JL, Doherty TM, Ratsitorahina M, Ramarokoto H, Zumla A, Huggett J, Rook G, Richard V, Gicquel B, Rasolofo-Razanamparany V; VACSEL/VACSIS Study Group. 2010. Variation in gamma interferon responses to different infecting strains of *Mycobacterium tuberculosis* in acid-fast bacillus smear-positive patients and household contacts in Antananarivo, Madagascar. *Clinical Vaccine Immunology* 17: 1094–1203 DOI:10.1128/CVI.00049-10
39. Wang C, Peyron P, Mestre O, Kaplan G, van Soolingen D, Gao Q, Gicquel B, Neyrolles O. 2010. Innate immune response to *Mycobacterium tuberculosis* Beijing and other genotypes. *PLoS ONE* 5: 1-8 DOI:10.1371/journal.pone.0013594
40. Couvin D, Reynaud Y, Rastogi N. 2019. Two tales: worldwide distribution of Central Asian (CAS) versus ancestral East-African Indian (EAI) lineages of *Mycobacterium tuberculosis* underlines a remarkable cleavage for phylogeographical, epidemiological and demographical characteristics. *PLoS ONE* 14: 1-20 doi.org/10.1371/journal.pone.0219706
41. Coker OO, Chaiprasert A, Ngamphiw C, Tongsima S, Regmi SM, Clark TG, Ong RT, Teo YY, Prammananan T, Palittapongarnpim P. 2016. Genetic signatures of *Mycobacterium tuberculosis* Nonthaburi genotype revealed by whole genome analysis of strains from tuberculous meningitis patients in Thailand. *PeerJ* 4: 1-20 DOI:10.7717/peerj.1905
42. Singh J, Sankar MM, Kumar P, Couvin D, Rastogi N, Singh S, Indian TB Diagnostics Network. 2015. Genetic diversity and drug susceptibility profile of *Mycobacterium tuberculosis* isolated from different regions of India. *Journal of Infectology* 71: 207–219 DOI:10.1016/j.jinf.2015.04.028
43. Rodríguez-Castillo JG, Pino C, Niño LF, Rozo JC, Llerena-Polo C, Parra-López CA, Tauch A, Murcia-Aranguren MI. 2017. Comparative genomic analysis of *Mycobacterium tuberculosis* Beijing-like strains revealed specific genetic variations associated with virulence and drug resistance. *Infection Genetics and Evolution* 54: 314–323 DOI:10.1016/j.meegid.2017.07.022

Figure 1

Phylogenetic analysis of *M. tuberculosis* strains from PTB and EPTB phenotypes of the disease.

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

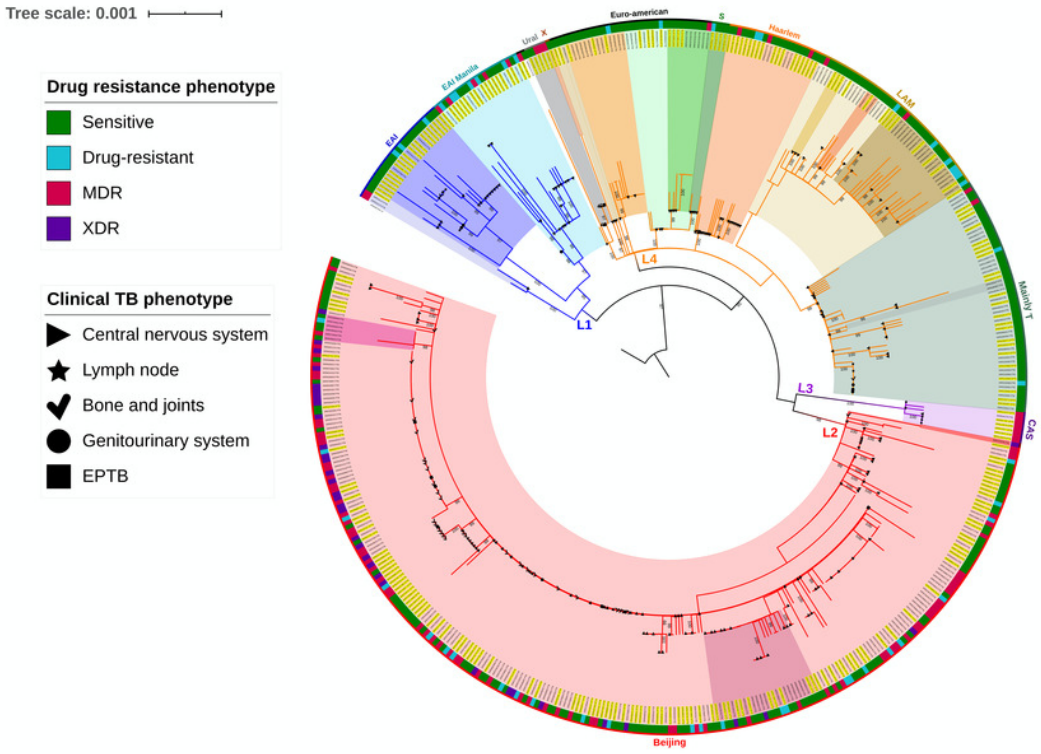

Figure 2

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

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