Tuberculosis xxx (2015) 1–7

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

EPIDEMIOLOGY

# Whole genome sequencing identifies circulating Beijing-lineage *Mycobacterium tuberculosis* strains in Guatemala and an associated urban outbreak

Joseph W. Saelens [a, 1], Dalia Lau-Bonilla [b, 1], Anneliese Moller [b], Narda Medina [b], Brenda Guzmán [b], Maylena Calderón [b], Raúl Herrera [b], Dana M. Sisk [a], Ana M. Xet-Mull [a], Jason E. Stout [c], Eduardo Arathoon [d, e], Blanca Samayoa [b, f, **], David M. Tobin [a, *]

[a] Department of Molecular Genetics and Microbiology and Center for Host-Microbial Interactions, Duke University Medical Center, Durham, NC, USA
[b] Asociación de Salud Integral, Guatemala City, Guatemala
[c] Department of Medicine, Division of Infectious Diseases and International Health, Duke University Medical Center, Durham, NC, USA
[d] Hospital General San Juan de Dios, Guatemala City, Guatemala
[e] Clínica Familiar Luis Ángel García, Guatemala City, Guatemala
[f] Facultad de Ciencias Químicas y Farmacia, Universidad de San Carlos de Guatemala, Guatemala City, Guatemala

ARTICLE INFO

*Article history:*
Received 16 July 2015
Received in revised form
22 August 2015
Accepted 5 September 2015

*Keywords:*
Tuberculosis
Molecular epidemiology
Whole genome sequencing
Beijing strains
Guatemala
Central America

SUMMARY

Limited data are available regarding the molecular epidemiology of *Mycobacterium tuberculosis* (*Mtb*) strains circulating in Guatemala. Beijing-lineage *Mtb* strains have gained prevalence worldwide and are associated with increased virulence and drug resistance, but there have been only a few cases reported in Central America. Here we report the first whole genome sequencing of Central American Beijing-lineage strains of *Mtb*. We find that multiple Beijing-lineage strains, derived from independent founding events, are currently circulating in Guatemala, but overall still represent a relatively small proportion of disease burden. Finally, we identify a specific Beijing-lineage outbreak centered on a poor neighborhood in Guatemala City.



## 1. Introduction

Tuberculosis (TB) poses a major challenge to global health. The causative agent, *Mycobacterium tuberculosis* (*Mtb*), was responsible for around 1.5 million deaths in 2013 [1]. The seven major lineages of *Mtb* have historically been associated with specific geographical regions, and substantial evidence exists that genetic differences between lineages influence disease presentation and outcome [2–6]. Beijing lineage strains, also known as Lineage 2 strains, have emerged as important drivers of global *Mtb* burden [7–9]. Notably, outside of East Asia, "modern" Beijing strains account for the majority of this burden [10]. These "modern" Beijing strains display elevated rates of drug-resistance, rapid progression of disease and increased transmission [11–16]. While there is still debate as to precisely when the Beijing lineage originated, it appears to have arisen in East Asia and, through successive waves of human migration, spread throughout this region and beyond [10,17]. Positive selection on virulence-associated genes as well as compensatory mutations negating the fitness-cost of drug-resistance are key factors in making Beijing the most successful contemporary lineage [17–20].

The emergence of "modern" Beijing strains to new regions has been particularly pronounced in Eastern Europe and Africa, regions in which Euro-American strains are also present [21–23]. South Africa serves as one striking example, where data suggest that Beijing strains have continually expanded since their arrival to account for an increasing proportion of total burden [20]. Beijing strains in South Africa have developed resistance to standard drug therapies and contributed to poor outcomes for patients [24,25].

* Corresponding author.
** Corresponding author.
*E-mail addresses:* investigacion@asi.org.gt (B. Samayoa), david.tobin@duke.edu (D.M. Tobin).
[1] These authors contributed equally.

http://dx.doi.org/10.1016/j.tube.2015.09.001

In the Western Hemisphere, the emergence of Beijing strains has thus far been more complex. Beijing strains have spread across the United States, with notable outbreaks occurring in New York City, NY, and Houston, TX [26–28]. In contrast, South and Central America have heretofore been thought to be relatively isolated from this trend, and Euro-American/Lineage 4 strains predominate [29–36]. Current evidence suggests Beijing strains have failed to gain a foothold in South America outside of Peru, where a substantial number of Chinese and Japanese immigrants settled throughout the 19th and 20th centuries [30]. However, there is very little information on *Mtb* strain diversity in Central America. Thus far, the sparse evidence that does exist indicates that Beijing strains contribute only minimally to the burden of tuberculosis in the region [17,37].

Here we investigate the status of Beijing-lineage strains in Guatemala, the most populous country in Central America. We take a whole genome sequencing (WGS) approach to assess recent TB cases at an HIV clinic and associated public hospital. We combine our results with global genome sequences as well as epidemiologic patient data to assess patterns of Beijing-lineage *Mtb* in Guatemala.

## 2. Materials and methods

### 2.1. Patient population

The Clínica Familiar "Luis Angel García" (CFLAG) is an HIV-specialized clinic located at the Hospital General San Juan de Dios, one of Guatemala's two public teaching hospitals in Guatemala City. It provides comprehensive treatment and follow-up for people living with HIV for both outpatient and inpatient care and has provided medical care for more than 10,000 patients over the last thirty years. From 2010 to 2014, routine spoligotyping was performed at CFLAG on all *Mtb* isolates from CFLAG and the associated hospital during that period, derived from a total of 514 independent patients. Spoligotyping identified 11 Beijing-lineage strains; five of the 11 strains could be regrown for WGS.

### 2.2. DNA extraction and genotyping

Bacterial culturing, spoligotyping, and DNA extraction were performed according to standard practices (detailed in Supplemental Material) [38,39].

### 2.3. Sequencing, alignment and SNP calling

Using Illumina HiSeq 2500 and Illumina MiSeq platforms, we sequenced each strain at >500-fold coverage with a minimum read length of 50 base pairs. Additional sequence reads acquired from published WGS datasets also had minimum read lengths of 50 base pairs. We downloaded reads from the National Center for Biotechnology Information Sequence Read Archive and the European Nucleotide Archive, as described in [2,3,17,40]. We collected sequence reads for 24 globally extant strains of *Mtb* representing the seven major lineages [2,3,41], as well as 96 globally extant Beijing strains [17]. We aligned reads against the H37Rv reference genome (GenBank: AL123456.3) using BWA [42]. We called variants using SAMtools and filtered with VarScan for a minimum read depth of 10, a consensus quality score of 20, and a minimum variant frequency of 0.75 [43,44]. We discarded SNPs adjacent to indels and within repetitive regions of the genome. Additionally, as is standard practice, we discarded SNPs affecting genes commonly associated with drug-resistance to avoid homoplastic mutations among distantly related strains [10,17,45]. We visualized variants among the Guatemalan isolates associated with the outbreak with Circular Visualization for Microbial Genomes (URL: http://civi.cmbi.ru.nl/).

### 2.4. Nucleotide sequence accession numbers

The sequences determined in this study have been deposited in the NCBI Sequence Read Archive under accession numbers: GG-135-10 – **SRR1765871**, GG-120-10 – **SRR1765872**, GG-152-12 – **SRR1765877**, GG-131-11 – **SRR1765874**, GG-219-11 – **SRR1765879**, and GG-30-13 – **SRR1765888**. GG-219-11 and GG-30-13 were sequential isolates taken from the same patient 17 months apart.

### 2.5. Tree builds

All trees were based on genome-wide SNPs derived according to the parameters specified above. We constructed a superset of SNPs for each strain with reference alleles occupying sites for which no variants were detected using custom Perl scripts. These SNPs informed neighbor-joining and maximum-likelihood methods of phylogeny construction. We implemented neighbor-joining methodology with ClustalW2 using pairwise similarity scores of SNP supersets as a measure of genetic distance [46]. We generated a maximum-likelihood phylogeny with RAxML using a GTR model of nucleotide substitution [47]. For each method, 1000 bootstrap replicates provided support for nodes on the tree. The phylogenies derived from each method were congruent (compare Figure 2 to Figure S1). Trees were visualized with FigTree (see URL tree.bio.ed.ac.uk/software/figtree).

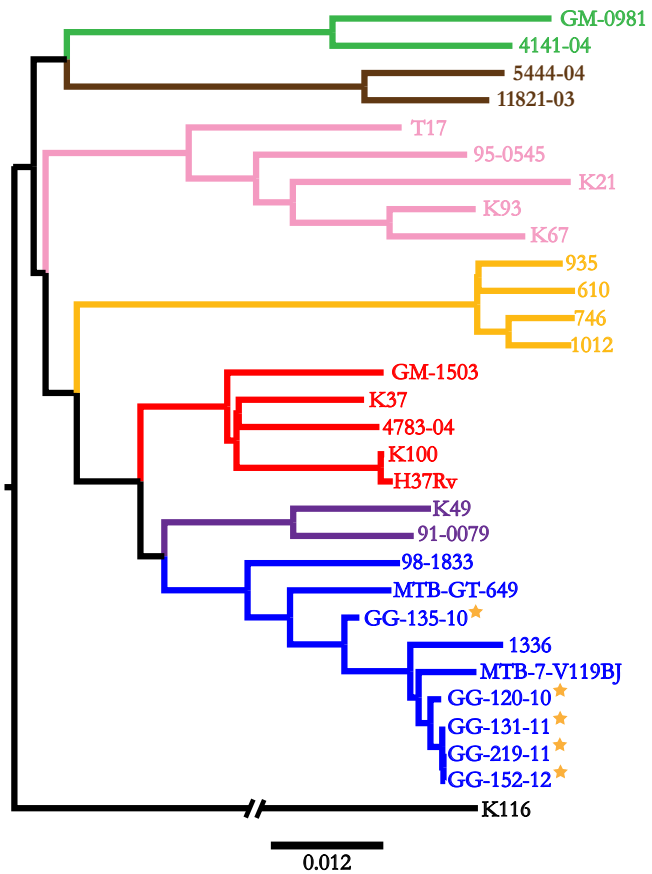


**Figure 1.** Neighbor-joining phylogeny based on 18,039 SNPs among 29 global strains of *Mtb*. The tree is rooted by the outgroup *M. canetti*. Branches are colored by lineage: Pink – Lineage 1/Indo-Oceanic; Blue – Lineage 2/East-Asian; Purple – Lineage 3/East-African-Indian; Red – Lineage 4/Euro-American; Brown – Lineage 5/West-Africa I; Green – Lineage 6/West Africa II; Yellow – Lineage 7/Ethiopia. Guatemalan isolates are marked with stars. Scale bar indicates substitutions per site.

Please cite this article in press as: Saelens JW, et al., Whole genome sequencing identifies circulating Beijing-lineage *Mycobacterium tuberculosis* strains in Guatemala and an associated urban outbreak, Tuberculosis (2015), http://dx.doi.org/10.1016/j.tube.2015.09.001