Contents lists available at ScienceDirect

International Journal of Medical Microbiology

journal homepage: www.elsevier.com/locate/ijmm

# Substantial molecular evolution and mutation rates in prolonged latent *Mycobacterium tuberculosis* infection in humans

Troels Lillebaek[a,*,1], Anders Norman[b,1], Erik Michael Rasmussen[a], Rasmus L. Marvig[b,2], Dorte Bek Folkvardsen[a], Åse Bengård Andersen[c], Lars Jelsbak[b,*]

[a] *International Reference Laboratory of Mycobacteriology, Statens Serum Institut, DK-2300 Copenhagen, Denmark*
[b] *Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark*
[c] *Department of Infectious Diseases, Copenhagen University Hospital, Rigshospitalet, DK-2100 Copenhagen, Denmark*

ARTICLE INFO

*Article history:*
Received 15 March 2016
Received in revised form 27 May 2016
Accepted 27 May 2016

*Keywords:*
Latent infection
Within-host evolution
Mutation rate
Genome sequencing
Pathogen evolution

ABSTRACT

The genome of *Mycobacterium tuberculosis* (Mtb) of latently infected individuals may hold the key to understanding the processes that lead to reactivation and progression to clinical disease. We report here analysis of pairs of Mtb isolates from putative prolonged latent TB cases. We identified two confirmed cases, and used whole genome sequencing to investigate the mutational processes that occur over decades in latent Mtb. We found an estimated mutation rate between 0.2 and 0.3 over 33 years, suggesting that latent Mtb accumulates mutations at rates similar to observations from cases of active disease.

## 1. Introduction

Failure to control tuberculosis (TB) is in part due to the ability of *Mycobacterium tuberculosis* (*Mtb*) to transition into a latent state, in which growth is curtailed by the host immune system and clinical signs of active disease are absent. People harboring latent *Mtb* have a 5–10% lifetime risk of contracting active TB (Dutta and Karakousis, 2014). The infection outcome (i.e. latent infection versus active disease) seemingly depends on complex interactions between host and bacterial factors as well as the overall immune- and health status of the patient (Taarnhoj et al., 2011). It is estimated that one-third of the global human population is infected with latent *Mtb*, constituting a vast disease reservoir. As a genetically monomorphic organism (Boritsch et al., 2014; Eldholm and Balloux, 2016), the adaptational potential of *Mtb, i.e.* its capacity to genetically adapt to the host immune system or drug interventions, is tightly linked to its *in vivo* mutation rate (Eldholm and Balloux, 2016; Ford et al., 2013; Takiff and Feo, 2015) and the type of accumulated mutations (synonymous *vs* non-synonymous). Despite this, dormancy remains an underreported aspect of *Mtb* biology, primarily due to limited availability of isolates collected both at the time of infection and at the onset of TB.

Denmark possess a unique *Mtb* culture collection comprising isolates from all culture positive TB cases in Denmark, Greenland and the Faeroe Islands from the last 24 years ($n = 9674$) and a historical collection of *Mtb* isolates sampled during the 1960s ($n = 203$). Using existing restriction fragment length typing (IS6110-RFLP) data, we have previously identified linked isolate-pairs separated by more than three decades, indicating cases of *Mtb* re-activation after prolonged latent infection (Lillebaek et al., 2002). In one particularly striking case, two such isolates were sampled from within the same household (Lillebaek et al., 2002), a father and his son, diagnosed with pulmonary tuberculosis in 1961 and 1994, respectively. By subjecting this and other similarly linked isolate-pairs to whole-genome sequencing, we are therefore presented with a rare and unique opportunity to shed more light on the mutational processes that occur over decades in latent *Mtb*.

## 2. Methods

### 2.1. Bacterial isolates

Historical isolates from the 1960s were originally stored as freeze-dried samples until they were re-cultured in 2001. All isolates were stored in 15% glycerol at −80 °C and were cultivated in

[*] Corresponding authors.
*E-mail addresses:* tll@ssi.dk (T. Lillebaek), lj@bio.dtu.dk (L. Jelsbak).
[1] These authors contributed equally.
[2] Present address: Center for Genomic Medicine, Rigshospitalet, 2100 Copenhagen, Denmark.

http://dx.doi.org/10.1016/j.ijmm.2016.05.017
1438-4221/© 2016 Elsevier GmbH. All rights reserved.

Dubos media prior to growth on blood agar. Two to three colonies were picked for DNA extraction, which were performed according to the procedure used for RFLP fingerprinting (van Soolingen et al., 1991).

### 2.2. Genome sequencing

All DNA samples were whole-genome sequenced on the Illumina HiSeq2000 platform as 2 × 100 bp paired-end libraries with 500 bp inserts, with the exception of Mu879, which was sequenced on the Illumina MiSeq platform as a 2 × 150 bp library (400 bp inserts) due to contamination of the initial DNA sample with *Staphylococcus aureus*. The sequence data described in this study is available from the European Nucleotide Archive (ENA) at this URL: http://www.ebi.ac.uk/ena/data/view/PRJEB10245.

### 2.3. Read processing and variant calling

Illumina paired-end reads were pre-trimmed, using the program Trimmomatic to completely remove any residual TruSeq adapter fragments as well as leading and trailing nucleotides with Phred quality-scores below 3 (Bolger et al., 2014). This was done to ensure optimal mapping conditions. Additionally, read-pairs in which one or both mates were shorter than 36 bp or had an average base quality below 15 were discarded prior to mapping. Remaining reads were mapped to the *Mtb* H37Rv (GenBank Accession no.: AL123456) or the *Mtb* CTRI-2 (Accession no.: CP002992) genome sequences, using the Burrows-Wheeler Alignment tool (BWA v0.7.10) (Li and Durbin, 2009). The Genome Analysis ToolKit (GATK) was used to accurately realign reads around long indels (DePristo et al., 2011). Median mapping depths ranged from 119 to 227 and average read lengths ranged from 96 to 150 bp (Table S1). In all 14 libraries >99% of reads mapped to the H37Rv reference genome, with >98% of its bases covered by at least one read. SAMtools v0.1.19 (Li et al., 2009) was used to call raw variants (using the mpileup and bcftools programs) from aligned reads with a minimum mapping quality (mapQ) of 30 to exclude reads mapping to multiple locations on the reference genome. Raw variants were then filtered so that only single nucleotide polymorphism (SNPs) covered by at least 5 reads and a minimum of one read in each direction. The minimum acceptable average mapQ of SNPs was set to 45. For the H37Rv reference genome only variants called as homozygous (0/0 or 1/1) in all samples by the bcftools variant caller were considered. For the CTRI-2 reference (Ilina et al., 2013), variant calling was less stringent, so that all variants supported by 85% or more of the mapped reads in a sample were seen as fixated, while variants with lower coverage were kept as transitory mutations, but otherwise not considered in final analyses. With the CTRI-2 reference, the four samples (Mu837, Mu838, R94-2977 and R93-3208), mapped reads had error rates ≤0.15%. Individual coverage of fixed SNPs ranged from 35 to 213. To resolve large deletions and SNPs around *Mtb* repetitive regions, scaffolds were assembled using the SPAdes genome assembler version 3.5.0 (http://bioinf.spbau.ru/spades) on quality-trimmed paired-end reads only. These were mapped onto the reference genome using the Geneious software's "Map to Reference" function (http://www.geneious.com). In resolving putative repetitive genes (such as *pe*-, *ppe*- or transposon genes) only regions in which repetitive elements were covered by scaffolds extending into non-repetitive regions by more than 1kbp were considered. Furthermore, all fixed SNPs called against the CTRI-2 reference were confirmed by visual inspection of mapped reads in Geneious.

An in-house perl script was written to combine genome annotations and variants from variant call format (vcf) files and to concatenate SNPs into alignments for phylogenetic analysis. Maximum likelihood phylogeny was calculated using PhyML (Guindon et al., 2010) with the GTR substitution model (-I -Γ) and Maximum parsimony was calculated using MEGA (Tamura et al., 2013). Bootstrap values were calculated from 500 trees to confirm strong branch support (>95%) for all *Mtb* sublineages as well as the phylogenetic placement of primary-, secondary- and tertiary isolates.

### 2.4. Variant effect analysis, IS-element mapping and dN/dS ratio estimates

The Protein Variation Effect Analyser (PROVEAN) software, which is a validated method for estimating the deleterious or neutral effects of individual amino acid variants (Choi et al., 2012), was used on intragenic SNPs. The program was run using default settings against the NCBI NR (non-redundant) protein database (version 2014-08-13). A Provean-score of −2.500, was used as an arbitrary threshold to distinguish between deleterious (below −2.5) and neutral (above −2.5) amino acid changes, which is slightly more conservative than the −2.182 used in the manuscript. The program IS-mapper (http://github.com/jhawkey/IS_mapper) was used to pin-point the location of insertions of the IS-element IS6110 along the CTRI-2 genome. The synonymous substitution rate per site (*dS*) to the non-synonymous substitution rate per site (*dN*) via the ratio $\omega = dN/dS$ was used as a global estimator of positive-, neutral- or purifying selection pressure. These were calculated by using the number of non-synonymous (2,690,224) and synonymous (979,340) sites in all non-repetitive coding regions of the *Mtb* H37Rv genome. Corresponding P-values were calculated under the assumption that the numbers of non-synonymous and synonymous mutations are independent Poisson random variables (Yang et al., 2011).

## 3. Results & discussion

### 3.1. Identification of true latent TB cases

To identify true cases of prolonged latent *Mtb* infection, a precise account of accumulated mutations occurring between infection and active disease is required. Re-activated isolates should harbor most if not all mutations observed in their putative origin, in addition to the mutations that have accumulated in the interim, and the primary isolate should ideally be placed as its most recent common ancestor (MRCA) on a phylogenetic tree. We therefore sequenced a total of 14 *Mtb* isolates from the two strain collections mentioned above. In total, 6 discrete isolate groups (Links 1–6), inferred from identical RFLP-profiles (Lillebaek et al., 2003), were evaluated. Each link comprised a primary "origin" isolate from the 1960s and a secondary (putatively re-activated) isolate from the 1990s. Additionally, Links 5 & 6 included tertiary isolates that were likely recent transmissions of the respective secondary isolates (Fig. 1a). Alignments of concatenated single nucleotide polymorphisms (SNPs) were used to determine the phylogeny of 14 isolates in relation to a global collection of sequenced reference *Mtb* isolates (Fig. 1b; Table S2). Links 1 & 2 were revealed to constitute a single monophyletic clade which, in addition to the previously described father and son isolates (Link 2; Mu838 and R94-3208), contained a primary isolate from the father's niece (Link 1; Mu837), diagnosed with TB in 1961, and a secondary isolate from an unknown individual from the same geographic region of Denmark (link 1; R94-2977), who contracted the disease in 1994. Given that Mu837 was isolated from a one-year-old child it is highly unlikely, although not impossible, that the child was the direct source for R94-2977 (the recipient was five years at the time) as small children are generally not considered to be very contagious with TB. However, even if this child, the only available case with an identical RFLP profile to R94-2977, was not the direct source of transmission, the difference of only a single SNP between Mu837 and the MRCA strongly indicates a close