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

Lillebaek, Troels; Norman, Anders; Rasmussen, Erik Michael; Marvig, Rasmus Lykke; Bek Folkvardsen, Dorte; Bengaard Andersen, Åse; Jelsbak, Lars

Published in:
International Journal of Medical Microbiology

Link to article, DOI:
10.1016/j.ijmm.2016.05.017

Publication date:
2016

Document Version
Peer reviewed version

Title: Substantial molecular evolution and mutation rates in prolonged latent Mycobacterium tuberculosis infection in humans

Author: Troels Lillebaek Anders Norman Erik Michael Rasmussen Rasmus L. Marvig Dorte Bek Folkvardsen Åse Bengård Andersen Lars Jelsbak

PII: S1438-4221(16)30110-2
DOI: http://dx.doi.org/doi:10.1016/j.ijmm.2016.05.017
Reference: IJMM 51063

To appear in:

Received date: 15-3-2016
Revised date: 27-5-2016
Accepted date: 27-5-2016

Please cite this article as: Lillebaek, Troels, Norman, Anders, Rasmussen, Erik Michael, Marvig, Rasmus L., Folkvardsen, Dorte Bek, Andersen, Åse Bengård, Jelsbak, Lars, Substantial molecular evolution and mutation rates in prolonged latent Mycobacterium tuberculosis infection in humans.International Journal of Medical Microbiology http://dx.doi.org/10.1016/j.ijmm.2016.05.017

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

Troels Lillebaek[1*#], Anders Norman[2#], Erik Michael Rasmussen[1], Rasmus L. Marvig[2§], Dorte Bek Folkvardsen[1], Åse Bengård Andersen[3], Lars Jelsbak[2*]

[1]International Reference Laboratory of Mycobacteriology, Statens Serum Institut, DK-2300 Copenhagen, Denmark.

[2]Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark.

[3]Department of Infectious Diseases, Copenhagen University Hospital, Rigshospitalet, DK-2100 Copenhagen, Denmark.

Running title*: M. tuberculosis* evolution in humans

# These authors contributed equally
§ Present address: Center for Genomic Medicine, Rigshospitalet, 2100 Copenhagen, Denmark.

*Correspondence to: Lars Jelsbak (lj@bio.dtu.dk) and Troels Lillebaek (tll@ssi.dk)

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. (Word-count: 93)

Keywords: *Mycobacterium tuberculosis*; latent infection; within-host evolution; mutation rate; genome sequencing; pathogen evolution

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*=9,674) 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*.

Methods

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 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).

Genome sequencing. All DNA samples were whole-genome sequenced on the Illumina HiSeq2000 platform as 2 x 100bp paired-end libraries with 500 bp inserts, with the exception of Mu879, which was sequenced on the Illumina MiSeq platform as a 2 x 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.

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 36bp 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 119-227 and average read lengths ranged from 96-150 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 35-213. 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.

Variant effect analysis, IS-element mapping and d_N/d_S 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 (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).

Results & Discussion

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 (**Figure 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 (**Figure 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 causal link between the two isolates. Therefore, it is very likely that transmission occurred within the same setting and within a relatively short time period (i.e. a few months) to both the one-year-old niece and the

latently infected individual. This is also supported by the complete lack of other isolates with the same typing profiles as R94-2977 in isolates collected since nationwide genotyping was introduced (in 1992) (Lillebaek et al., 2003), which is unlikely to have occurred in strains that were actively transmitted between 1961 and 1992. The difference of 5 SNPs and one IS6110-element (**Table 1**; **Table S5**) between Link 1 & 2 MRCAs also makes the father's isolate (Mu838) a highly unlikely transmission origin given that they were both diagnosed with TB in late 1961.

None of the other putative linked isolates (Links 3-6) placed the primary isolate sufficiently close (<3 SNPs) to the MRCA of the secondary- and third isolates, to confirm re-activation. These cases therefore seem to be parts of broader and largely undocumented chains of transmission. In our view, however, this does not rule out the possibility that some of these secondary isolates underwent prolonged latent infection. Both of the tertiary isolates displayed only single additional SNPs in relation to their respective putative predecessors (the two reactivated isolates), confirming their status as recent transmissions.

Molecular clock rates. Due to the rather limited number of mutations observed in Links 1 and 2 (**Table S3**), we performed a second round of variant calling on these four isolates, in which a more closely related *Mtb* strain, CTRI-2, was used as a reference (**Figure 1c**; **Table 1**). The four isolates displayed 97 common SNPs and 17 indels in relation to CTRI-2 (**Table S4**). We also used *de novo* assembled scaffolds to ensure the validity of all SNP-calls inside or in the proximity of known repetitive regions, which led to the identification of three additional SNPs. Link 1 showed the accumulation of eight SNPs and one indel over 33 years, while Link 2 (the father and the son) showed the accumulation of 10 SNPs during the same time-period. The two links (i.e. the two primary isolates) were themselves separated by five SNPs. Thus, we find an average molecular clock rate of 0.2 – 0.3 SNPs per genome per year. Interestingly, this compares quite well to rates of mutation acquisition observed previously in Lineage 4, both *in*

vitro (Ford et al., 2011; Ford et al., 2013) and in several independent retrospective outbreak studies (Eldholm et al., 2015; Guerra-Assuncao et al., 2015; Walker et al., 2013). While current models of *Mtb* during latency assume that there is little or no growth or mutation of the infecting bacteria, a recent *in vitro* study using a macaque model has provided evidence that *Mtb* mutates at a fixed rate over time, and that mutation rates are comparable between latent and active disease (Ford et al., 2011). However, while these and other results (Ford et al., 2013) suggest that the mutation rate remains largely constant irrespective of disease state, it is important to emphasize that such observations have so far been characterized by small sample sizes. Furthermore, Colangeli *et al* (Colangeli et al., 2014) have provided contrasting evidence by showing mutation rates ten times lower during latency than active disease by examining four strains, two of them latent for more than 20 years, derived from the same index case.

Mutational patterns and within-host selection The diversifying processes that contribute to mutation accumulation during long-term latent infection remain poorly understood. Such processes can be investigated by studying mutational patterns as revealed by genome sequencing. We first examined a previously proposed hypothesis that mutation accumulation during latency is often the product of oxidative damage (Ford et al., 2011). Links 1 and 2 were inspected for molecular signatures indicative of oxidative damage, either cytosine deamination (GC to AT changes) or formation of 8-oxoguanine (GC to TA changes). For Link 2, 30% of observed SNPs were GC>AT or GC>TA mutations, highly comparable to the 36% observed in the CTRI-2 background. Despite Link 1 having a notably higher proportion of these mutations (75%), it was not quite statistically significant (Fisher's exact test, $P = 0.054$) due to the low number of observed polymorphisms. We therefore found no real evidence supporting that oxidative stress is the dominant contributing factor during latent infection in our data. A similar observation was also made by Colangeli *et al* (Colangeli et al., 2014).

We also investigated the role of within-host evolution during latent infection by comparing synonymous and non-synonymous polymorphisms in the two only confirmed re-activated isolates as well as the number of putatively deleterious mutations. A previous study revealed significantly reduced levels of purifying selection (average pairwise dN/dS = 0.57) throughout the Mycobacterium tuberculosis complex (MTBC) with typically 40% of amino acid changes having functional consequences (Hershberg et al., 2008). A similar ratio (0.53) was also observed prior to a major *Mtb* outreak (pre-diversification) among Inuit in the Nunavik region of Canada, while significantly elevated ratios (0.95 and 0.76, respectively) were observed during the outbreak in two separate sub-lineages (post-diversification)(Lee et al., 2015). Thus, the ratio observed in the CTRI-2 background and Link 1 (dN/dS = 0.56) corresponds well with Nunavik pre-diversification rates and those observed overall throughout the MTBC. Interestingly, we found a significant overrepresentation of both non-synonymous SNPs (dN/dS = 2.5; Poisson, P = 0.0011) and putatively deleterious mutations (86%; Fisher's exact test; P = 0.0032) in Link 2, demonstrating a clear divergence from the CTRI-2 background. The underlying cause of this variation in the mutational pattern is unknown, and could merely reflect random genetic drift combined with serial bottlenecking owing to low effective population sizes. Such factors are influenced by distinct environmental conditions at different sites of infection or lesion types in the infected individual. Analysis of multiple isolates from latent infections is therefore required to fully understand the role of the different mutational processes and selective forces acting on the latent bacteria.

Gene selection. We did not observe mutations in any of the secondary- or tertiary isolates that could be linked to established drug resistance phenotypes. We note, however, that genes Rv3805c (*aftB*) and Rv3130c (*tgs1*), which both contained putative deleterious mutations in isolate R98-3208 (Link 2), have been previously implicated in processes that may influence the development of antibiotic resistance and tolerance

(Daniel et al., 2004; Safi et al., 2013). However, the putatively deleterious *tgs1* mutation in R98-3208 is particularly intriguing in a purely metabolic context as it is essential for accumulation of triacylglycerol (TAG) in response to hypoxia and other stresses (Daniel et al., 2004). Furthermore, TAG-synthesis plays a direct role in promoting entry into a dormancy-like state (Daniel et al., 2004). Increased expression of *tgs1,* observed in certain Beijing lineages of *Mtb,* is speculated to offer increased survival within host tissues as a consequence of high TAG-levels (Reed et al., 2007).

Concluding remarks

Overall, our study underpins the challenges associated with identification of relevant bacterial isolates, and highlights the importance of systematic sampling and whole genome sequencing of *Mtb* in relation to refining our understanding of underlying mechanisms of long-term persistence in the host. While these and other observations contribute to our understanding of the bacterial *in vivo* lifecycle it remains to be seen if and how these findings scale over prolonged latent TB infections in humans. The rate of molecular evolution observed here in combination with the finding that mutations with relevance for the physiological state of the infecting bacteria may emerge during latency, emphasize the importance of identifying and controlling cases of latent infections.

Word-count: 2790

Acknowledgements

This work was supported by the Novo Nordisk Foundation (grant number NNF7651) and the Lundbeck Foundation (grant number R151-2013-14628). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

We thank Pia Kristiansen for patient and skillful laboratory work retrieving the historical strains from old collections.

Conflicts of interest

None of the authors have either a commercial or other association that might pose a conflict of interest.

References

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Boritsch, E.C., Supply, P., Honore, N., Seemann, T., Stinear, T.P., Brosch, R., 2014. A glimpse into the past and predictions for the future: the molecular evolution of the tuberculosis agent. Mol Microbiol 93, 835-852.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., Chan, A.P., 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS One 7, e46688.

Colangeli, R., Arcus, V.L., Cursons, R.T., Ruthe, A., Karalus, N., Coley, K., Manning, S.D., Kim, S., Marchiano, E., Alland, D., 2014. Whole genome sequencing of Mycobacterium tuberculosis reveals slow growth and low mutation rates during latent infections in humans. PLoS One 9, e91024.

Daniel, J., Deb, C., Dubey, V.S., Sirakova, T.D., Abomoelak, B., Morbidoni, H.R., Kolattukudy, P.E., 2004. Induction of a novel class of diacylglycerol acyltransferases and triacylglycerol accumulation in Mycobacterium tuberculosis as it goes into a dormancy-like state in culture. Journal of bacteriology 186, 5017-5030.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491-498.

Dutta, N.K., Karakousis, P.C., 2014. Latent tuberculosis infection: myths, models, and molecular mechanisms. Microbiology and molecular biology reviews : MMBR 78, 343-371.

Eldholm, V., Balloux, F., 2016. Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out. Trends Microbiol.

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., Balloux, F., 2015. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. Nat Commun 6, 7119.

Ford, C.B., Lin, P.L., Chase, M.R., Shah, R.R., Iartchouk, O., Galagan, J., Mohaideen, N., Ioerger, T.R., Sacchettini, J.C., Lipsitch, M., Flynn, J.L., Fortune, S.M., 2011. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. Nat Genet 43, 482-486.

Ford, C.B., Shah, R.R., Maeda, M.K., Gagneux, S., Murray, M.B., Cohen, T., Johnston, J.C., Gardy, J., Lipsitch, M., Fortune, S.M., 2013. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet 45, 784-790.

Guerra-Assuncao, J.A., Crampin, A.C., Houben, R.M., Mzembe, T., Mallard, K., Coll, F., Khan, P., Banda, L., Chiwaya, A., Pereira, R.P., McNerney, R., Fine, P.E., Parkhill, J., Clark, T.G., Glynn, J.R., 2015. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife 4.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology 59, 307-321.

Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A., Feldman, M.W., Gagneux, S., 2008. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS biology 6, e311.

Ilina, E.N., Shitikov, E.A., Ikryannikova, L.N., Alekseev, D.G., Kamashev, D.E., Malakhova, M.V., Parfenova, T.V., Afanas'ev, M.V., Ischenko, D.S., Bazaleev, N.A., Smirnova, T.G., Larionova, E.E., Chernousova, L.N., Beletsky, A.V., Mardanov, A.V., Ravin, N.V., Skryabin, K.G., Govorun, V.M., 2013. Comparative genomic analysis of Mycobacterium tuberculosis drug resistant strains from Russia. PLoS One 8, e56577.

Lee, R.S., Radomski, N., Proulx, J.F., Levade, I., Shapiro, B.J., McIntosh, F., Soualhine, H., Menzies, D., Behr, M.A., 2015. Population genomics of Mycobacterium tuberculosis in the Inuit. Proc Natl Acad Sci U S A 112, 13609-13614.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Genome Project Data Processing, S., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Lillebaek, T., Dirksen, A., Baess, I., Strunge, B., Thomsen, V.O., Andersen, A.B., 2002. Molecular evidence of endogenous reactivation of Mycobacterium tuberculosis after 33 years of latent infection. The Journal of infectious diseases 185, 401-404.

Lillebaek, T., Dirksen, A., Vynnycky, E., Baess, I., Thomsen, V.O., Andersen, A.B., 2003. Stability of DNA patterns and evidence of Mycobacterium tuberculosis reactivation occurring decades after the initial infection. The Journal of infectious diseases 188, 1032-1039.

Reed, M.B., Gagneux, S., Deriemer, K., Small, P.M., Barry, C.E., 3rd, 2007. The W-Beijing lineage of Mycobacterium tuberculosis overproduces triglycerides and has the DosR dormancy regulon constitutively upregulated. Journal of bacteriology 189, 2583-2589.

Safi, H., Lingaraju, S., Amin, A., Kim, S., Jones, M., Holmes, M., McNeil, M., Peterson, S.N., Chatterjee, D., Fleischmann, R., Alland, D., 2013. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-beta-D-arabinose biosynthetic and utilization pathway genes. Nat Genet 45, 1190-1197.

Taarnhoj, G.A., Engsig, F.N., Ravn, P., Johansen, I.S., Larsen, C.S., Roge, B., Andersen, A.B., Obel, N., 2011. Incidence, risk factors and mortality of tuberculosis in Danish HIV patients 1995-2007. BMC Pulm Med 11, 26.

Takiff, H.E., Feo, O., 2015. Clinical value of whole-genome sequencing of Mycobacterium tuberculosis. Lancet Infect Dis 15, 1077-1090.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular biology and evolution 30, 2725-2729.

van Soolingen, D., Hermans, P.W., de Haas, P.E., Soll, D.R., van Embden, J.D., 1991. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. Journal of clinical microbiology 29, 2578-2586.

Walker, T.M., Ip, C.L., Harrell, R.H., Evans, J.T., Kapatai, G., Dedicoat, M.J., Eyre, D.W., Wilson, D.J., Hawkey, P.M., Crook, D.W., Parkhill, J., Harris, D., Walker, A.S., Bowden, R., Monk, P., Smith, E.G., Peto, T.E., 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13, 137-146.

Yang, L., Jelsbak, L., Marvig, R.L., Damkiaer, S., Workman, C.T., Rau, M.H., Hansen, S.K., Folkesson, A., Johansen, H.K., Ciofu, O., Hoiby, N., Sommer, M.O., Molin, S., 2011. Evolutionary dynamics of bacteria in a human host environment. Proc Natl Acad Sci U S A 108, 7481-7486.

Figure legends

Fig. 1 A proposed timeline (a) of routes of transmission (dotted arrows) between *Mtb* isolates from 14 TB patients based on existing IS6110-RFLP typing. Each line begins at birth and ends with the onset of active TB (filled circles) followed by the patients' age (in years) at the time. Lighter areas represent periods in which the patient is assumed to have been uninfected, while darkly colored areas is the assumed period of (latent) TB. A Maximum-likelihood analysis (b) based on 9,401 single-nucleotide polymorphisms (SNPs) showed that all belong to the Euro-American *Mtb* lineage (Lineage 4). Sublineage assignments are based on presence of established informative SNPs and lineage-defining regions of difference (RDs). A maximum-parsimony analysis (c) of the four *M. tuberculosis* clones from Links 1 & 2, based on SNPs relative to *Mtb* CTRI-2 (Lineage 4.3.3 reference). Individual branches are labeled in roman numerals, followed by the number of SNPs on that branch in brackets. Year of isolation are indicated in brackets at each branch tip.

Table 1. Single nucleotide polymorphisms (SNPs) of *Mtb* isolates from Links 1 & 2 called using *Mtb* CTRI-2 as the reference genome.

Position	Variant	Gene	SNP Frequencies[a]				ORF NT change	Synonymous SNP	AA Change[b]	Annotation[c]
			Mu837	R94-2977	Mu838	R94-3208				
						Branch I[d]				
3,902,730	G->A	mce4F (MTCTRI2_3560)	0.02 (61)	0.00 (106)	**0.98 (63)**	**1.00 (37)**	C1530T	N	P510L	MCE-family protein MCE4F
						Branch II				
250,770	C->T	MTCTRI2_0214	0.00 (117)	0.00 (136)	0.00 (119)	**1.00 (109)**	C895T	N	**Q299***	Hypothetical protein
842,514	A->T	.	0.00 (138)	0.00 (192)	0.00 (102)	**1.00 (136)**	.	.	.	Intergenic region (MTCTRI2_0766/MTCTRI2_0767)
1,102,021	T->C	.	0.00 (133)	0.00 (188)	0.00 (129)	**0.99 (129)**	.	.	.	Intergenic region (mscL/MTCTRI2_1010)
1,291,604	G->C	narH (MTCTRI2_1194)	0.00 (146)	0.00 (187)	0.00 (116)	**1.00 (116)**	G177C	N	**Q59H**	Respiratory nitrate reductase subunit beta NarH
1,569,291	T->C	cyp132 (MTCTRI2_1431)	0.00 (128)	0.00 (148)	0.00 (125)	**1.00 (119)**	A621G	N	Y207C	Cytochrome P450 132
2,461,641	G->A	ctaC (MTCTRI2_2235)	0.00 (122)	0.00 (192)	0.00 (81)	**1.00 (90)**	C188T	N	**P63S**	Transmembrane cytochrome C oxidase subunit II CtaC
3,492,216	A->G	tgs1 (MTCTRI2_3193)	0.00 (137)	0.00 (243)	0.00 (122)	**0.99 (137)**	T1296C	N	**L432P**	Triacylglycerol synthase
4,046,395	C->G	ppe65 (MTCTRI2_3686)	0.00 (117)	0.00 (186)	0.00 (117)	**1.00 (93)**	G1072C	Y	.	**PPE family protein**
4,182,139	G->A	ctpJ (MTCTRI2_3816)	0.00 (116)	0.00 (152)	0.00 (127)	**1.00 (104)**	C225T	N	**A75V**	Cation transporter P-type ATPase CtpJ
4,255,227	A->G	aftB (MTCTRI2_3884)	0.00 (108)	0.00 (164)	0.00 (137)	**1.00 (127)**	T600C	N	**M200T**	Possible arabinofuranosyltransferase AftB
						Branch III				
1,862,175	T->A	argR (MTCTRI2_1686)	**1.00 (117)**	**1.00 (164)**	0.00 (119)	0.00 (96)	T83A	N	**I28N**	Arginine repressor

1,862,176	C->T	argR (MTCTRI2_1686)	**1.00 (116)**	**1.00 (163)**	0.00 (118)	0.00 (97)	C84T	N	.	Arginine repressor
2,337,938	G->A	.	**1.00 (37)**	**1.00 (63)**	0.00 (32)	0.00 (36)	.	.	.	Intergenic region (lppJ/MTCTRI2_2120)
2,337,943	G->A	.	**1.00 (35)**	**1.00 (62)**	0.00 (27)	0.00 (33)	.	.	.	Intergenic region (lppJ/MTCTRI2_2120)
2,837,047	G->A	fas (MTCTRI2_2572)	**0.99 (86)**	**1.00 (130)**	0.02 (63)	0.00 (77)	C7191T	N	S2397L	Fatty acid synthase
					Branch IV					
1,274,252	C->A	mmpL13b (MTCTRI2_1178)	**1.00 (121)**	0.00 (158)	0.00 (110)	0.00 (119)	C536A	N	**A179E**	Transmembrane transport protein MmpL13B
					Branch V					
404,025	C->A	MTCTRI2_0343	0.00 (99)	**1.00 (153)**	0.00 (92)	0.00 (81)	C1487A	N	P496H	**13E12 repeat family protein**
534,952	G->A	ufaA1 (MTCTRI2_0450)	0.00 (121)	**1.00 (179)**	0.00 (132)	0.00 (133)	C1277T	N	P426S	Cyclopropane-fatty-acyl-phospholipid synthase
534,953	G->A	ufaA1 (MTCTRI2_0450)	0.00 (121)	**1.00 (172)**	0.00 (129)	0.00 (134)	C1276T	Y	.	Cyclopropane-fatty-acyl-phospholipid synthase
2,432,547	G->A	idsA2 (MTCTRI2_2208)	0.00 (110)	**1.00 (177)**	0.00 (118)	0.00 (140)	G568A	N	V190I	Geranylgeranyl pyrophosphate synthetase
2,751,892	G->A	MTCTRI2_2500	0.00 (126)	**1.00 (207)**	0.00 (127)	0.00 (133)	C1124T	Y	.	Oxidoreductase alpha subunit
3,116,605	C->G	MTCTRI2_2870	0.00 (79)	**0.98 (112)**	0.00 (78)	0.00 (3)	G284C	N	**A95P**	**Transposase**
3,669,145	>C<	pcd (MTCTRI2_3360)	0.00 (142)	**1.00 (197)**	0.00 (159)	0.00 (135)	.	.	.	Piperideine-6-carboxylic acid dehydrogenase
4,271,165	C->G	MTCTRI2_3898	0.00 (151)	**1.00 (236)**	0.00 (139)	0.00 (176)	C181G	N	**R61G**	Hypothetical protein
4,361,068	C->T	esxD (MTCTRI2_3970)	0.00 (155)	**0.99 (159)**	0.00 (119)	0.00 (131)	G295A	Y	.	ESAT-6 like protein EsxD

[a] SNP frequencies are followed by mapping depth in parenthesis. Alleles with a frequency above 0.85 were considered fixated (bold).

[b] Amino acid changes marked in bold are putatively deleterious based on Provean analysis.

[c] Gene annotations marked in bold are normally filtered out *in Mtb* variant analysis but were validated by visual inspection of mapped reads in this study.

[d] Branch numbers refer to branches in Fig. 1c