Reference set of Mycobacterium tuberculosis clinical strains: A tool for research and product development

Journal Article

Author(s):
Borrell, Sònia; Trauner, Andrej; Brites, Daniela; Rigouts, Leen; Loiseau, Chloe; Coscolla, Mireia; Niemann, Stefan; De Jong, Bouke; Yeboah-Manu, Dorothy; Kato-Maeda, Midori; Feldmann, Julia; Reinhardt, Miriam; Beisel, Christian; Gagneux, Sebastien

Publication date:
2019

Permanent link:
https://doi.org/10.3929/ethz-b-000336262

Originally published in:
PLoS ONE 14(3), https://doi.org/10.1371/journal.pone.0214088

PLOS ONE

RESEARCH ARTICLE

Reference set of *Mycobacterium tuberculosis* clinical strains: A tool for research and product development

Sònia Borrell[1,2]*, Andrej Trauner[1,2], Daniela Brites[1,2], Leen Rigouts[3,4], Chloe Loiseau[1,2], Mireia Coscolla[1,2], Stefan Niemann[5], Bouke De Jong[3], Dorothy Yeboah-Manu[6], Midori Kato-Maeda[7], Julia Feldmann[1,2], Miriam Reinhard[1,2], Christian Beisel[8], Sebastien Gagneux[1,2]

1 Swiss Tropical and Public Health Institute, Basel, Switzerland, **2** University of Basel, Basel, Switzerland, **3** Mycobacteriology Unit, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium, **4** Collection of Mycobacterial Cultures (BCCM/ITM), Institute of Tropical Medicine, Antwerp, Belgium, **5** Division of Molecular and Experimental Mycobacteriology Group, Research Center Borstel, Borstel, Germany, **6** Noguchi Memorial Institute for Medical Research, University of Ghana, Accra, Ghana, **7** School of Medicine, University of California at San Francisco, San Francisco, California, United States of America, **8** Genomics Facility, Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

* sonia.borrell@swisstph.com

Citation: Borrell S, Trauner A, Brites D, Rigouts L, Loiseau C, Coscolla M, et al. (2019) Reference set of *Mycobacterium tuberculosis* clinical strains: A tool for research and product development. PLoS ONE 14(3): e0214088. https://doi.org/10.1371/journal.pone.0214088

Editor: Olivier Neyrolles, Institut de Pharmacologie et de Biologie Structurale, FRANCE

Received: October 22, 2018

Accepted: March 6, 2019

Published: March 25, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This work was supported by the Swiss National Science Foundation (grants 310030_166687, IZRJZ3_164171, IZLSZ3_170834 and CRSII5_177163), and by the European Research Council (309540-EVODRTB), and SystemsX.ch.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The *Mycobacterium tuberculosis* complex (MTBC) causes tuberculosis (TB) in humans and various other mammals. The human-adapted members of the MTBC comprise seven phylogenetic lineages that differ in their geographical distribution. There is growing evidence that this phylogeographic diversity modulates the outcome of TB infection and disease. For decades, TB research and development has focused on the two canonical MTBC laboratory strains H37Rv and Erdman, both of which belong to Lineage 4. Relying on only a few laboratory-adapted strains can be misleading as study results might not be directly transferrable to clinical settings where patients are infected with a diverse array of strains, including drug-resistant variants. Here, we argue for the need to expand TB research and development by incorporating the phylogenetic diversity of the MTBC. To facilitate such work, we have assembled a group of 20 genetically well-characterized clinical strains representing the seven known human-adapted MTBC lineages. With the "MTBC clinical strains reference set" we aim to provide a standardized resource for the TB community. We hope it will enable more direct comparisons between studies that explore the physiology of MTBC beyond the laboratory strains used thus far. We anticipate that detailed phenotypic analyses of this reference strain set will increase our understanding of TB biology and assist in the development of new control tools that are broadly effective.

Introduction

Tuberculosis (TB) remains an urgent public health problem causing 10.4 million new cases and 1.3 million deaths every year [1]. The TB epidemic is worsening due to growing drug resistance and the absence of a universally effective vaccine against the transmissible pulmonary form of the disease [1].

The outcome of TB infection and disease is highly variable, ranging from rapid clearing by the innate immune response to life-long latent infection and various forms of active pulmonary and extra-pulmonary disease. In the past, most of this variation was attributed to the host and environmental factors. Because of the limited genetic diversity within the *Mycobacterium tuberculosis* complex (MTBC) compared to other bacteria [2], the view has been that no relevant phenotypic variation should be expected. However, recent advances in whole genome sequencing of large MTBC clinical strain collections from global sources have revealed more genomic diversity than previously appreciated. Specifically, the human-adapted MTBC comprises seven phylogenetic lineages that differ in their geographic distribution, and individual members of these lineages can differ by up to ~2,000 single nucleotide polymorphisms (SNPs). This is equivalent to the phylogenetic distance between *M*. *tuberculosis* sensu stricto and *M*. *bovis*, which is a typical pathogen of cattle.

In addition to the genomic diversity across MTBC clinical strains, findings from many experimental studies have led to a change in paradigm by demonstrating the phenotypic impact of this genetic diversity. For example, studies have reported differences between clinical strains with respect to their transcriptomic profiles [3, 4], protein and metabolite levels [5] [4], methylation profiles [5], drug susceptibility [6] and cell wall structure [7–9]. In addition, MTBC genetic diversity has also been shown to influence disease severity and human to human transmission, with "modern" lineages showing a faster progression to disease and shorter latency periods compared to strains from the "ancestral" clades [10–13].

Most of what we know about TB biology today is based on work performed during many decades, most of which has relied on the two canonical laboratory strains H37Rv and Erdman. Both of these strains, as well as the clinical strain CDC1551 used by some TB laboratories more recently, belong to MTBC Lineage 4 [14]. A notable exception is HN878, which belongs to Lineage 2 and is a gaining prominence as a laboratory representative of the Beijing family of strains [15].

H37Rv was first isolated from a patient (H37) with pulmonary tuberculosis in 1905 at the Trudeau Sanatorium in Saranac Lake, New York, while Erdman was isolated from human sputum by William H. Feldman in 1945, at Mayo Clinic, Rochester.

Since its original isolation, H37Rv has been used extensively in biomedical research. The sequence of its genome was published by Cole and colleagues in 1998, which was a breakthrough in TB research [16]. Indeed, H37Rv and its genome sequence still provide the backbone for most of TB biological research today, informing studies ranging from basic biochemistry and microbiology to global omics profiling, systems biology, drug discovery and immunology. However, H37Rv has been passaged countless times in various laboratories, and despite retaining its virulence in mice, it has adapted to laboratory conditions [17]. The same is likely true for Erdman and CDC1551 which have been isolated later than H37Rv, but which by now, have also been passaged in the laboratory for several decades. Hence, despite the great progress in our understanding of TB generated through studies based on laboratory strains, there are good reasons to expect that the findings from these studies do not paint the full picture and could benefit from being validated in more genetic backgrounds.

Despite the increasing number of experimental studies revealing important phenotypic differences across MTBC clinical strains, many of these studies have been difficult to reproduce

between different laboratories, and the data are often contradictory. Moreover, linking experimental phenotypes to clinical and epidemiological characteristics of MTBC lineages or strains has been particularly challenging. We propose that part of these challenges could be overcome by standardizing the complement of clinical MTBC strains we study. As a first step, we suggest to broaden the scope of basic and translational TB research by incorporating a set of genetically well-characterized clinical strains representative of the known phylogenetic diversity of the pathogen. In time, the community would accumulate a significant body of data that could support new findings that are more relevant to global TB. To this end it is important that there is a collective agreement to avoid passaging these strains extensively and minimize laboratory adaptation.

Over the years, our group has been collecting strains from around the world and characterizing them by whole genome sequencing. Our main aim was to draw evolutionary and phylogeographic inferences [18], however, we also realized the importance of studying this diversity more broadly, which is why we used our global collection of MTBC clinical strains and the associated phylogenomic data to rationally select a subset to be used as reference strains for future research. We believe this set of strains will be of value for the TB research community.

The "MTBC clinical strain reference set" comprises 20 clinical strains covering all 7 known human-adapted MTBC lineages. These strains have been submitted to the Mycobacterial culture bank of the Belgian Coordinated Collections of Microorganism (BCCM/ITM) and will be available for anyone interested in the phenotypic impact of MTBC diversity (http://bccm.belspo.be/).

Material and methods

Strain selection

We based our initial selection of strains to be included in this reference set on phylogenetic trees that were built with a combination of genomes from our collection and other publicly available genomes representing the known global diversity of the human-adapted MTBC [19]. Initially, we picked 43 strains that were intended to represent a diverse sampling of each lineage, comprising several sub-lineages where appropriate. We strove to include strains that represent as much as possible the phylogenetic breadth of each lineage, thus attempting to capture most of the within-lineage diversity The strains had to be free of known drug resistance mutations and carry only genomic deletions that were congruous with their phylogenetic background, without any rare genomic abnormalities. Moreover, we included strains that were already used in experimental work in the past; N0072, N0157, N0031, N0145 and N0155 [4, 20]. The remaining strains were selected from the large number of isolates present in the combined collections of the authors. Specifically: N0004, N0054, N0069 and N0136 were contributed by UCSF, University of California; N1268, N1272, N1274 and N1283, were contributed by the Research Center in Borstel, Germany; N1176, N1201, N1202 and N1216 were contributed by the Noguchi Memorial Institute for Medical Research in Accra, Ghana; N0091 was contributed by MRC-Gambia and N3913 in the Victorian Infectious Diseases Reference Laboratory in Melbourne. None of the strains were isolated specifically for this study.

Bacterial culture and DNA extraction

All MTBC isolates included into the "MTBC clinical strain reference set", were processed and derived from single colonies. Strains were grown in 7H9/Tween 0.05% medium (BD) +/- 40mM sodium pyruvate. We extracted genomic DNA for whole genome sequencing (WGS) from cultures in the late exponential phase of growth using the CTAB method [21].

Phenotypical drug susceptibility test (DST)

DST was performed for the main anti-TB drugs by the proportion method, using the following drug concentrations: RMP (4 μg/ml), INH (0.2 and 1.0 μg/ml), EMB (2.0 μg/ml) and SM (4.0 μg/ml) on Löwenstein-Jensen medium, and OFX (2.0 and 8.0 μg/ml), KAN (6 μg/ml), CAP (10 μg/ml) and ETH (10 μg/ml) on Middlebrook 7H11 agar.

Spoligotyping

Spoligotyping was performed according to internationally standardized protocols [22]. We used KvarQ to derive *in silico* spoligotypes from FASTQ files containing the WGS information [23] when necessary.

Whole-genome sequencing

Sequencing libraries were prepared using NEXTERA XT DNA Preparation Kit (Illumina, San Diego, USA). Multiplexed libraries were paired-end sequenced on Illumina HiSeq2500 (Illumina, San Diego, USA) with 151 or 101 cycles at the Genomics Facility Basel.

Sequence read alignment and variant determination

The obtained FASTQ files were processed with Trimmomatic v 0.33 (SLIDINGWINDOW: 5:20) [24] to clip Illumina adaptors and trim low quality reads. Any reads shorter than 20 bp were excluded for the downstream analysis. Overlapping paired-end reads were then merged with SeqPrep v 1.2 (overlap size = 15) (https://github.com/jstjohn/SeqPrep). We used BWA v 0.7.13 (mem algorithm) [25] to align the resultant reads to the reconstructed ancestral sequence of MTBC obtained in [19]. Duplicated reads were marked by the Mark Duplicates module of Picard v 2.9.1 (https://github.com/broadinstitute/picard) and excluded. The Realigner Target Creator and Indel Realigner modules of GATK v 3.4.0 [26] were used to perform local realignment of reads around indels. To avoid false positive calls Pysam v 0.9.0 (https://github.com/pysam-developers/pysam) was used to exclude reads with alignment score lower than (0.93*read_length)-(read_length*4*0.07)), corresponding to more than 7 miss-matches per 100 bp. SNPs were called with Samtools v 1.2 mpileup [27] and VarScan v 2.4.1 [28] using the following thresholds: minimum mapping quality of 20, minimum base quality at a position of 20, minimum read depth at a position of 7-fold and without strand bias. Only SNPs considered to have reached fixation within a patient were considered (at a within-host frequency of $\geq$90%). Conversely, when the SNP within-host frequency was $\leq$10% the ancestor state was called. Additionally, we excluded genomes with average coverage $<$ 15-fold (after all the referred filtering steps). All SNPs were annotated using snpEff v4.11, in accordance with the *M. tuberculosis* H37Rv reference annotation (NC000962). SNPs falling in regions such as PPE and PE-PGRS, phages, insertion sequences and in regions with at least 50 bp identities to other regions in the genome were excluded from the analysis as in [29]. Drug resistance-conferring mutations were annotated based on a previously published list [23]. Determination of sub-lineage was done using the phylogenetic SNPs according to Stucki *et al.* [29] and to Coll *et al.* [30].

Detection of genomic duplications

We used the output of Samtools v1.2 mpileup and VarScan 2.4.1 to extract the mapping coverage depth of short sequencing reads (coverage) per genomic position. We split the genome into bins of 500 base pairs and calculated the median coverage for each bin. We then computed the z-score for all the bins across the genome of each strain. Finally, we calculated the median

of coverage z-score medians for 80 consecutive bins spanning 40,000 base pairs. We plotted both the z-scores and z-score pool medians and identified large genomic deletions by visual inspection of the resultant coverage plots. Duplicated regions appear as regions with a marked increase in the apparent coverage.

Phylogenetic reconstruction

To put the reference strain set into a phylogenetic context, we combined them into a phylogeny containing publicly available genomes (n = 232) [19]. We used all 16,614 variable positions to infer a Maximum Likelihood phylogeny using the MPI parallel version of RAxML [31]. We used the GTR model as implemented in RAxMLto perform 1,000 rapid bootstrap inferences, followed by a thorough maximum-likelihood search [31]. We show the best-scoring Maximum Likelihood topology. The phylogeny was rooted using *Mycobacterium canettii* as an out-group.

Results and discussion

Selection of the "MTBC clinical strain reference set"

Starting from the 43 candidates, we excluded strains that we could not re-grow in the laboratory and those that had any mutation known to confer drug resistance [32]. We then selected 20 pan-susceptible clinical strains, based on full genome data for inclusion as reference strains (Table 1). The selected set covers all 7 known phylogenetic lineages of the human-adapted MTBC, but does not include any animal-adapted members of the MTBC. For most lineages, we selected 3 representative strains to maximize the within- and between-lineage diversity (Fig 1). Given the current limited availability of Lineage 7 strains, we were able to include only a

Table 1. "MTBC clinical strains reference set" list.

Strain	Place of Birth	Lin	Spoligotypea	Date of isolation	Sub-Lin	+Pyr	DR	SIT	BCCM/ITM n°
N0069	China	L1	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■■□□□	9/11/1998	L1.1.1		wt	nd	ITM-2018-00102
N0072	India	L1	■□□■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■□□□□■■■	6/15/1997	L1.1.2		wt	355	ITM-2018-00083
N0157	Philippines	L1	■■□■■■■■■■■■■■■■■□□■■■■■■■□□□□■□■■■■■■■■■	6/15/1999	L1.2.1		wt	19	ITM-2018-00082
N0031	China	L2	■■■	6/15/1994	L2.1		wt	523	ITM-2018-00087
N0052	China	L2	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■	6/15/1998	L2.2.2		wt	1	ITM-2018-02241
N0145	China	L2	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■	6/15/1997	L2.2.1.1		wt	1	ITM-2018-00085
N0155	China	L2	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■	6/15/1998	L2.2.1		wt	1	ITM-2018-00088
N0004	India	L3	■■■□□□□■■■■■■■■■■□■■■■□□□□□□□□□□□□□□■■■■■■■	3/1/2010	nd		wt	nd	ITM-2018-00089
N0054	Ethiopia	L3	■■■□□□□■■■■■■■■■■■■■■□□□□□□□□□□□□■■□□■■■■■	6/15/1999	nd		wt	309	ITM-2018-00091
N1274	Afghanistan	L3	■■■□□□□■■■■■■■■■■■■■■□□□□□□□□□□□□□□□■■■■■■	2/15/2005	nd		STR	357	ITM-2018-00090
N0136	USA	L4	■■■■■■■■■■■■■■■■■□□□□□■■■■■■■□■■■□□□□■■■■■■■	8/19/1991	L4.3.3		wt	222	ITM-2018-00093
N1216	Ghana	L4	■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■■■□□□□■■■■■■■	2/15/2011	L4.6.2.2		wt	61	ITM-2018-00092
N1283	Germany	L4	■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□■□□□□■■■■■■■	5/15/2003	L4.2.1		wt	35	ITM-2018-00094
N1176	Ghana	L5	■■■■■■■□□□□□■■■■■■■■□□□□■■■■■■■■■■■□□□■■■■	9/11/2009	nd	yes	wt	331	ITM-2018-00095
N1268	Sierra Leone	L5	■■■■■■■□□□□□■■■■■■□□□□□□■■■■■■■■■■■□□□■■■■	12/15/2003	nd	yes	wt	761	ITM-2018-00097
N1272	Ghana	L5	■■■■■■■□□□□□■■■■■■■□□□□■■■■■■■■■■■□□□□■■■■	6/1/2002	nd	yes	wt	330	ITM-2018-00096
N0091	Gambia	L6	■■■■■■□□□■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■	6/14/2002	nd	yes	wt	181	ITM-2018-00099
N1201	Ghana	L6	■■■■■■□□□■■■■■■■■■■□□□■■■■■■■■■■■■■□□■■■	2/17/2009	nd	yes	wt	nd	ITM-2018-00098
N1202	Ghana	L6	■□■■■■□□□■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■	2/5/2008	nd	yes	wt	318	ITM-2018-00100
N3913	Ethiopia	L7	■■■□□□□□□□□□□□□□□□□□□□□■■■□□■■■■■■■■■■■■■	11/6/2011	nd		wt	910	ITM-2018-00101

Place of birth, genotyping data (SNP typing and splogotyping[a]), date of isolation, sublineage classification based on Coll et al. [30], suggested growing conditions, SIT number and BCCM/ITM number for the strain bank.

[a]Spoligotyping data for each strain are shown, where black squares indicate the presence of a particular spacer and a white square the absence of a particular spacer.

https://doi.org/10.1371/journal.pone.0214088.t001

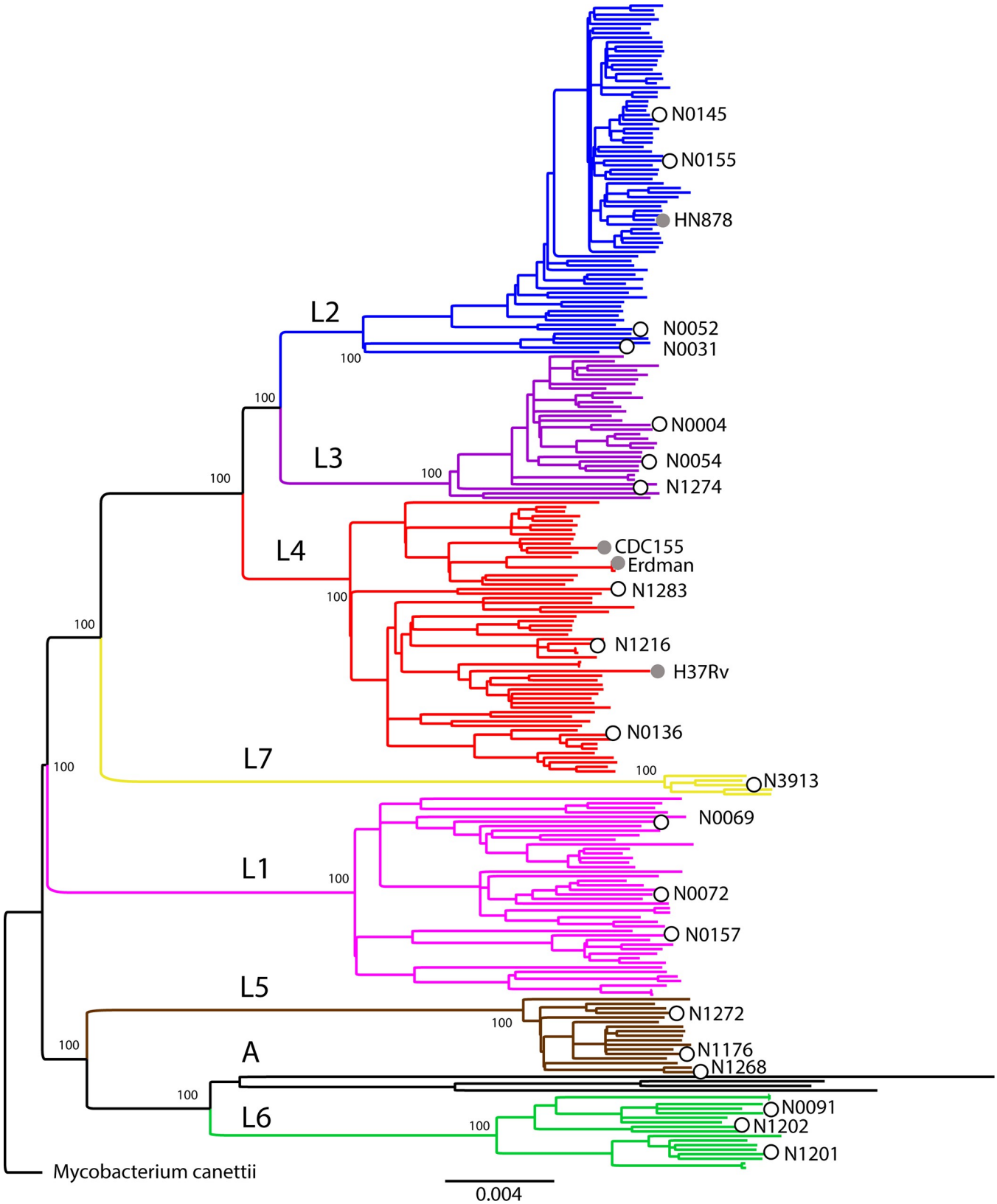

Fig 1. Maximum Likelihood topology of the 20 reference strains (open circles) plus 236 genomes representative of MTBC global diversity. Branch lengths are proportional to nucleotide substitutions and the topology is rooted with *Mycobacterium canettii*. Bootstrap values for clades corresponding to main MTBC lineages are shown. Grey circles indicate the phylogenetic placement of laboratory *M. tuberculosis* strains commonly used. "A" stands for animal MTBC.

https://doi.org/10.1371/journal.pone.0214088.g001

single representative. In the case of Lineage 2, we chose 4 strains to cover the wide range of genomic deletions found in this lineage [33]. These include N0155, a clinical strain that was used by our group in the past [20], and all the Lineage 2 strains that were transcriptionally profiled by Rose et al. [4]. We also included a "Proto-Beijing" strain (N0031), which belongs to a Lineage 2 clade that is phylogenetically basal compared to all other Lineage 2/Beijing strains. Proto-Beijing strains are also characterized by a deletion of RD105 but no deletion in RD207, which differentiates "proto-Beijing" from the regular "Beijing strains [34]. Moreover, N0031 shows an ancestral spoligotype with all DR spaces present (Table 1). Its basal branching provides an important contrast to the classical Beijing strains which carry deletions in both RD105 and RD207 [34, 35]. The Lineage 1 strains N0157 and N0072 have also been transcriptionally profiled [4]. In the case of Lineage 4, we selected representative strains of the "generalist" and "specialist" groups as defined by Stucki *et al* [29], respectively N0136 and N1216. The Lineage 4 strain N1283 is a representative of the sub-lineage L4.2 [29], which is also referred to as "Ural" based on spoligotyping [36]. The phylogenetic relationships of the 20 reference strains with respect to other representative MTBC strains are shown in Fig 1. Phenotypic resistance to STR was observed in N1274. This strain did not carry any mutation in the most common STR DR associated genes; rpsL or in the region 530_900 of rrs, however, a rare mutation was found in gydB (Rv3919c) at the AA position R137P. The rest of the "MTBC clinical strains reference set" was confirmed to be phenotypically drug susceptible to all main TB drugs.

Genomic characteristics

All annotated SNPs and insertions/deletions (indels) identified by comparison with the reconstructed MTBC ancestor sequence [37] and considered fixed (at frequency of $\geq$90%) for each strain are provided as supplementary files (S1 and S2 Tables). The general characteristics of the genome of each strain are presented in Table 2. We were able to observe all the large genomic deletions reported before [33] as gaps in sequencing coverage. For example, all Lineage 2 strains carried the deletion in RD105 and all but the Proto-Beijing strain also had a deletion in RD207. N0145 and N0155 shared the deletion in RD181, while N0145 harboured an additional deletion in RD150.

Given several reports in the literature regarding the importance of the duplication of a part of the genome that includes DosR and DosS (Rv3133c and Rv3134c) for MTBC virulence [4], we looked for areas of excessive read coverage within the genomes (S1 Fig). We identified four strains showing evidence of overlapping duplications covering DosR/S–N0031, N0145, N0155 and N1283 (S2 Fig). The first three strains belong to Lineage 2 while N1283 belongs to Lineage 4, corroborating past suggestions of convergent evolution [38, 39]. We did not detect any other genomic duplications of a comparable size in the genomes (S1 Fig).

Recommendations for growing and preserving the reference strains

Strains have been deposited in the Belgian Coordinated Collections of Microorganism (BCCM) and can be obtained from the BCCM/ITM: http://bccm.belspo.be/about-us/bccm-itm. Upon receipt, we suggest to grow a large culture of each strain in 7H9 (BD) and freeze multiple glycerol stocks for future use to avoid the acquisition of genetic changes due to laboratory adaptation during sequential sub-culturing [17]. Note that some strains require the addition of 40mM sodium pyruvate for optimal growth (Table 1).

Conclusions

For decades, TB research has almost exclusively focused on the two laboratory-adapted MTBC reference strains H37Rv and Erdman. Both strains have provided a common language across

Table 2. Characteristics of the "MTBC clinical strains reference set" genomes.

Strain	Coverage[a]	SNPs[b]	Indels[b]	% Genome Covered[c]	AC_Number[d]
N0069	81.09	898	135	98.11	ERR2704679
N0072	72.11	894	129	98.31	ERR2704680
N0157	74.32	894	132	98.87	ERR2704704
					ERR2704685
N0031	66.8	845	104	98.57	ERR2704676
N0052	110.37	862	93	98.98	ERR2704677
					ERR2704699
					ERR2704698
N0145	39.46	875	95	98.84	ERR2704702
					ERR2704701
					ERR2704683
N0155	115.19	897	105	99.14	ERR2704703
					ERR2704684
N0004	46.34	873	102	98.98	ERR2704675
					ERR2704696
					ERR2704697
N0054	64.21	886	110	98.4	ERR2704678
N1274	80.82	874	111	98.25	ERR2704693
N0136	52.92	823	52	99.15	ERR2704682
					ERR2704700
N1216	66.75	817	52	98.93	ERR2704705
					ERR2704689
N1283	52.97	831	61	98.97	ERR2704709
					ERR2704694
N1176	76.08	934	146	98.37	ERR2704686
N1268	51.11	937	134	98.58	ERR2704706
					ERR2704690
N1272	73.57	908	141	98.39	ERR2704708
					ERR2704707
					ERR2704692
					ERR2704691
N0091	72.87	1049	147	98.36	ERR2704681
N1201	77.64	1055	148	98.39	ERR2704687
N1202	78.02	1015	144	98.25	ERR2704688
N3913	100.62	1021	149	99.02	ERR2704711
					ERR2704695
					ERR2704710

[a] Average read depth after mapping and filtering out duplicated reads.

[b] Number of SNPs and short Indels considered fixed.

[c] Percentage of the reference chromosome (H37Rv) to which reads have been mapped.

[d] Accession Run Number.

https://doi.org/10.1371/journal.pone.0214088.t002

TB laboratories allowing knowledge to be built incrementally, with interoperable protocols, results and resources. However, insufficient attention has been given to the fact that both of these strains show patterns of laboratory adaptation and that they do not adequately represent the phylogenetical breadth of the human-adapted MTBC.

The new "MTBC clinical reference set" presented here covers much of this diversity and will provide the TB research community the opportunity to go beyond one single strain/lineage. The potential of sharing and integrating the experimental data generated with this strain set will enrich our understanding of the relationship between genotype and phenotype and potentially lead to fundamental new insights into TB biology. The true impact of genetic diversity in MTBC is slowly coming into focus; however there are still considerable gaps in our understanding. For example, it is known that clinical isolates show variations in drug susceptibility, but the basis for this is unclear [40]. Similarly, the association between drug resistance and specific strain backgrounds has been proposed in several studies, but the underlying mechanism remains unknown [9]. Vaccine and diagnostics development are two areas where understanding the impact of genetic diversity could be key to delivering effective products [14]. Similarly, we are only beginning to scratch the surface of the interplay between bacterial and human genetics at the immune interface [41]. These aspects of MTBC physiology deserve further attention especially due to their potential to have real clinical relevance. At a minimum, testing new TB diagnostics, drugs and vaccines against this strain set will help ensure these innovations are broadly effective.

Supporting information

S1 Table. "MTBC clinical strains reference set" SNPs list.
(ZIP)

S2 Table. "MTBC clinical strains reference set" Indels list.
(ZIP)

S1 Fig. "MTBC clinical strains reference set" with large genomic duplications.
(DOCX)

S2 Fig. Genome duplications affecting *dosR/dosS* overlap across strains.
(DOCX)

Acknowledgments

Computation was performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

Author Contributions

Conceptualization: Sònia Borrell, Andrej Trauner, Daniela Brites, Sebastien Gagneux.

Data curation: Sònia Borrell, Daniela Brites, Sebastien Gagneux.

Formal analysis: Sònia Borrell, Daniela Brites, Chloe Loiseau, Mireia Coscolla.

Funding acquisition: Sebastien Gagneux.

Investigation: Sònia Borrell, Sebastien Gagneux.

Methodology: Sònia Borrell, Andrej Trauner, Leen Rigouts, Chloe Loiseau, Julia Feldmann, Miriam Reinhard, Christian Beisel.

Project administration: Sònia Borrell, Julia Feldmann, Miriam Reinhard, Sebastien Gagneux.

Resources: Sònia Borrell, Stefan Niemann, Bouke De Jong, Dorothy Yeboah-Manu, Midori Kato-Maeda, Miriam Reinhard, Christian Beisel, Sebastien Gagneux.

Software: Sònia Borrell, Sebastien Gagneux.

Supervision: Sònia Borrell, Andrej Trauner, Sebastien Gagneux.

Validation: Miriam Reinhard, Christian Beisel, Sebastien Gagneux.

Writing – original draft: Sònia Borrell, Sebastien Gagneux.

Writing – review & editing: Sònia Borrell, Andrej Trauner, Daniela Brites, Leen Rigouts, Chloe Loiseau, Mireia Coscolla, Sebastien Gagneux.

References

1. WHO. World Health Organization. Global tuberculosis control—surveillance, planning, financing. (WHO, Geneva, Switzerland, 2017). 2017.
2. Achtman M., Evolution population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol. 2008; 62:53–70. https://doi.org/10.1146/annurev.micro.62.081307.162832 PMID: 18785837
3. Homolka S, Niemann S, Russell DG, Rohde KH. Functional genetic diversity among Mycobacterium tuberculosis complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. PLoS Pathog. 2010; 6(7):e1000988. https://doi.org/10.1371/journal.ppat.1000988 PMID: 20628579
4. Rose G, Cortes T, Comas I, Coscolla M, Gagneux S, Young DB. Mapping of genotype-phenotype diversity among clinical isolates of mycobacterium tuberculosis by sequence-based transcriptional profiling. Genome Biol Evol. 2013; 5(10):1849–62. https://doi.org/10.1093/gbe/evt138 PMID: 24115728
5. Zhu L, Zhong J, Jia X, Liu G, Kang Y, Dong M, et al. Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. Nucleic Acids Res. 2016; 44(2):730–43. https://doi.org/10.1093/nar/gkv1498 PMID: 26704977
6. Rouse DA, Morris SL. Molecular mechanisms of isoniazid resistance in Mycobacterium tuberculosis and Mycobacterium bovis. Infect Immun. 1995; 63(4):1427–33. PMID: 7890405
7. Portevin D, Sukumar S, Coscolla M, Shui G, Li B, Guan XL, et al. Lipidomics and genomics of Mycobacterium tuberculosis reveal lineage-specific trends in mycolic acid biosynthesis. Microbiologyopen. 2014; 3(6):823–35. https://doi.org/10.1002/mbo3.193 PMID: 25238051
8. Constant P, Perez E, Malaga W, Laneelle MA, Saurel O, Daffe M, et al. Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the Mycobacterium tuberculosis complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene. J Biol Chem. 2002; 277(41):38148–58. https://doi.org/10.1074/jbc.M206538200 PMID: 12138124
9. Coscolla M. Biological and Epidemiological Consequences of MTBC Diversity. Adv Exp Med Biol. 2017; 1019:95–116. https://doi.org/10.1007/978-3-319-64371-7_5 PMID: 29116631
10. Stavrum R, PrayGod G, Range N, Faurholt-Jepsen D, Jeremiah K, Faurholt-Jepsen M, et al. Increased level of acute phase reactants in patients infected with modern Mycobacterium tuberculosis genotypes in Mwanza, Tanzania. BMC Infect Dis. 2014; 14:309. https://doi.org/10.1186/1471-2334-14-309 PMID: 24903071
11. de Jong BC, Hill PC, Aiken A, Jeffries DJ, Onipede A, Small PM, et al. Clinical presentation and outcome of tuberculosis patients infected by M. africanum versus M. tuberculosis. Int J Tuberc Lung Dis. 2007; 11(4):450–6. PMID: 17394693
12. Portevin D, Gagneux S, Comas I, Young D. Human macrophage responses to clinical isolates from the Mycobacterium tuberculosis complex discriminate between ancient and modern lineages. PLoS Pathog. 2011; 7(3):e1001307. https://doi.org/10.1371/journal.ppat.1001307 PMID: 21408618
13. Krishnan N, Malaga W, Constant P, Caws M, Tran TH, Salmons J, et al. Mycobacterium tuberculosis lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. PLoS One. 2011; 6(9):e23870. https://doi.org/10.1371/journal.pone.0023870 PMID: 21931620
14. Gagneux S, Small PM. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infect Dis. 2007; 7(5):328–37. https://doi.org/10.1016/S1473-3099(07)70108-1 PMID: 17448936
15. Manca C, Tsenova L, Barry CE 3rd, Bergtold A, Freeman S, Haslett PA, et al. Mycobacterium tuberculosis CDC1551 induces a more vigorous host response in vivo and in vitro, but is not more virulent than other clinical isolates. J Immunol. 1999; 162(11):6740–6. PMID: 10352293

16. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998; 393(6685):537–44. https://doi.org/10.1038/31159 PMID: 9634230

17. Ioerger TR, Feng Y, Ganesula K, Chen X, Dobos KM, Fortune S, et al. Variation among genome sequences of H37Rv strains of Mycobacterium tuberculosis from multiple laboratories. J Bacteriol. 2010; 192(14):3645–53. https://doi.org/10.1128/JB.00166-10 PMID: 20472797

18. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol. 2018; 16 (4):202–13. https://doi.org/10.1038/nrmicro.2018.8 PMID: 29456241

19. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. 2013; 45(10):1176–82. https://doi.org/10.1038/ng.2744 PMID: 23995134

20. Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, Bohannan BJ. The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. Science. 2006; 312(5782):1944–6. https://doi.org/10.1126/science.1124410 PMID: 16809538

21. Parish TaS N.G. Isolation of genomic DNA from Mycobacteria. Mycobacteria Protocols–Methods in Molecular Biology ed. 1998; Totowa, New Jersey: Humana Press Inc:pp. 31–44.

22. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol. 1997; 35(4):907–14. PMID: 9157152

23. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics. 2014; 15:881. https://doi.org/10.1186/1471-2164-15-881 PMID: 25297886

24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

26. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

27. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27(21):2987–93. https://doi.org/10.1093/bioinformatics/btr509 PMID: 21903627

28. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22(3):568–76. https://doi.org/10.1101/gr.129684.111 PMID: 22300766

29. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet. 2016; 48 (12):1535–43. https://doi.org/10.1038/ng.3704 PMID: 27798628

30. Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 2014; 5:4812. https://doi.org/10.1038/ncomms5812 PMID: 25176035

31. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22(21):2688–90. https://doi.org/10.1093/bioinformatics/btl446 PMID: 16928733

32. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015; 15(10):1193–202. https://doi.org/10.1016/S1473-3099(15)00062-6 PMID: 26116186

33. Tsolaki AG, Gagneux S, Pym AS, Goguet de la Salmoniere YO, Kreiswirth BN, Van Soolingen D, et al. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis. J Clin Microbiol. 2005; 43(7):3185–91. https://doi.org/10.1128/JCM.43.7.3185-3191.2005 PMID: 16000433

34. Flores L, Van T, Narayanan S, DeRiemer K, Kato-Maeda M, Gagneux S. Large sequence polymorphisms classify Mycobacterium tuberculosis strains with ancestral spoligotyping patterns. J Clin Microbiol. 2007; 45(10):3393–5. https://doi.org/10.1128/JCM.00828-07 PMID: 17699643

35. Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, et al. Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese. Proc Natl Acad Sci U S A. 2015; 112(26):8136–41. https://doi.org/10.1073/pnas.1424063112 PMID: 26080405

36. Mokrousov I. The quiet and controversial: Ural family of Mycobacterium tuberculosis. Infect Genet Evol. 2012; 12(4):619–29. https://doi.org/10.1016/j.meegid.2011.09.026 PMID: 22036706

37. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet. 2010; 42(6):498–503. https://doi.org/10.1038/ng.590 PMID: 20495566

38. Weiner B, Gomez J, Victor TC, Warren RM, Sloutsky A, Plikaytis BB, et al. Independent large scale duplications in multiple M. tuberculosis lineages overlapping the same genomic region. PLoS One. 2012; 7(2):e26038. https://doi.org/10.1371/journal.pone.0026038 PMID: 22347359

39. Domenech P, Rog A, Moolji JU, Radomski N, Fallow A, Leon-Solis L, et al. Origins of a 350-kilobase genomic duplication in Mycobacterium tuberculosis and its impact on virulence. Infect Immun. 2014; 82 (7):2902–12. https://doi.org/10.1128/IAI.01791-14 PMID: 24778110

40. Colangeli R, Jedrey H, Kim S, Connell R, Ma S, Chippada Venkata UD, et al. Bacterial Factors That Predict Relapse after Tuberculosis Therapy. The New England journal of medicine. 2018; 379(9):823–33. https://doi.org/10.1056/NEJMoa1715849 PMID: 30157391

41. Brites D, Gagneux S. The Nature and Evolution of Genomic Diversity in the Mycobacterium tuberculosis Complex. Adv Exp Med Biol. 2017; 1019:1–26. https://doi.org/10.1007/978-3-319-64371-7_1 PMID: 29116627